Methods in Ecology and Evolution

Coupling validation effort with *in situ* bioacoustic data improves estimating relative activity and occupancy for multiple species with cross-species misclassifications

| Abstract: | 1. The increasing complexity and pace of ecological change requires natural resource managers to consider entire species assemblages. Acoustic recording units (ARUs) require minimal cost and effort to deploy and inform relative activity, or encounter rates, for multiple species simultaneously. ARU-based surveys require post-processing of the recordings via software algorithms that assign a species label to each recording. The automated classification process can result in cross-species misidentifications that should be accounted for when employing statistical modeling for conservation decision-making.

2. Using simulation and ARU-based detection counts from 17 bat species in British Columbia, Canada, we investigate three strategies for adjusting statistical inference for species misclassification: 1) ``coupling" ambiguous and unambiguous detections by validating a subset of survey events post-hoc, 2) using a calibration data set on the software algorithm's (in)accuracy for species identification, or 3) specifying informative Bayesian priors on classification probabilities. We explore the impact of different Bayesian prior specifications for the classification probabilities on posterior estimation. We then consider how the quantity of data validated post-hoc impacts model convergence and resulting inferences for bat species relative activity as related to nightly conditions and yearly site occupancy after accounting for site-level environmental variables.

3. Coupled methods resulted in less bias and uncertainty when |

estimating relative activity and species classification probabilities relative to calibration approaches. We found that species that were difficult-to-detect and those that were often inaccurately identified by the software required more validation effort than more easily detected and/or identified species.

4. Our results suggest that, when possible, acoustic surveys should rely on coupled validated detection information to account for false-positive detections, rather than uncoupled calibration data sets. However, if the assemblage of interest contains a large number of rarely detected or less prevalent species, an intractable amount of effort may be required, suggesting there are benefits to curating a calibration data set that is representative of the observation process. Our findings provide insights into the practical challenges associated with statistical analyses of ARU data and possible analytical solutions to support reliable and cost-effective decision-making for wildlife conservation/management in the face of known sources of observation errors.

Coupling validation effort with *in situ* bioacoustic data improves estimating relative activity and occupancy for multiple species with cross-species misclassifications

Christian A. Stratton[1], Kathryn M. Irvine[2], Katharine M. Banner[1], Wilson J. Wright[3], Cori Lausen[4], Jason Rae[5]

[1] *Montana State University, Bozeman, Montana, USA*

[2] *U.S. Geological Survey, Northern Rocky Mountain Science Center, Bozeman, Montana, USA*

[3] *Colorado State University, Fort Collins, Colorado, USA*

[4] *Wildlife Conservation Society Canada, Kalso, British Columbia, Canada*

[5] *Wildlife Conservation Society Canada, Nelson, British Columbia, Canada*

Corresponding author: Kathryn M. Irvine, email: `kirvine@usgs.gov`

Abstract: 345 words

Main text: 6078 words

1

Abstract

1. The increasing complexity and pace of ecological change requires natural resource managers to consider entire species assemblages. Acoustic recording units (ARUs) require minimal cost and effort to deploy and inform relative activity, or encounter rates, for multiple species simultaneously. ARU-based surveys require post-processing of the recordings via software algorithms that assign a species label to each recording. The automated classification process can result in cross-species misidentifications that should be accounted for when employing statistical modeling for conservation decision-making.

2. Using simulation and ARU-based detection counts from 17 bat species in British Columbia, Canada, we investigate three strategies for adjusting statistical inference for species misclassification: 1) "coupling" ambiguous and unambiguous detections by validating a subset of survey events post-hoc, 2) using a calibration data set on the software algorithm's (in)accuracy for species identification, or 3) specifying informative Bayesian priors on classification probabilities. We explore the impact of different Bayesian prior specifications for the classification probabilities on posterior estimation. We then consider how the quantity of data validated post-hoc impacts model convergence and resulting inferences for bat species relative activity as related to nightly conditions and yearly site occupancy after accounting for site-level environmental variables.

3. Coupled methods resulted in less bias and uncertainty when estimating relative activity and species classification probabilities relative to calibration approaches. We found that species that were difficult-to-detect and those that were often inaccurately identified by the software required more validation effort than more

2

easily detected and/or identified species.

4. Our results suggest that, when possible, acoustic surveys should rely on coupled validated detection information to account for false-positive detections, rather than uncoupled calibration data sets. However, if the assemblage of interest contains a large number of rarely detected or less prevalent species, an intractable amount of effort may be required, suggesting there are benefits to curating a calibration data set that is representative of the observation process. Our findings provide insights into the practical challenges associated with statistical analyses of ARU data and possible analytical solutions to support reliable and cost-effective decision-making for wildlife conservation/management in the face of known sources of observation errors.

1   Introduction

2   The increasing complexity and pace of ecological change requires natural resource man-

3   agers to consider entire species assemblages as opposed to the historic single-species focus

4   on understanding biological responses to management actions, climate change, and other

5   emerging threats (Baumgardt et al., 2018; Morganti et al., 2019; Reichert et al., 2021).

6   Efficient data collection methods are fundamental to achieving sufficient sample sizes for

7   adequately estimating complex, potentially interactive, species-environment relationships

8   across range-wide extents. Acoustic recording units (ARUs) require minimal cost and effort

9   to deploy and facilitate quantifying measures of activity for multiple species simultaneously.

10 Additionally, ARUs are effective at monitoring cryptic species, do not require invasive setup

11 procedures, and can be easily implemented by volunteers (Newson et al., 2017; Beason et al.,

12 2019). The culmination of these factors has encouraged programs to include acoustic data

13 as a primary source in large-scale monitoring efforts for many taxonomic groups, including

14 birds, bats and frogs (Loeb et al., 2015; Measey et al., 2017; Shonfield and Bayne, 2017;

15 Reichert et al., 2021).

16   While ARUs provide natural resource monitoring programs with a cost-effective means

17 for collecting multi-species data, the hidden expense is the post-processing of recordings

18 as the species labels automatically assigned to each recording via software algorithms are

19 not always correct. The automated classification process results in cross-species misiden-

20 tifications, leading to both false-positive and false-negative detections (Gibb et al., 2019;

21 Wright et al., 2020). Occupancy models provide a natural framework for statistical anal-

22 ysis of acoustic data, as they inherently account for false-negative detections. However,

4

23 standard single-species occupancy models assume that all false-positive detections are re-

24 moved prior to analysis (MacKenzie et al., 2002). Consequently, analysis of ARU data with

25 standard occupancy models requires additional confirmation of species presence during a

26 survey interval at each location, which can be cost-prohibitive for large-scale monitoring

27 efforts (Chambert et al., 2015; Guillera-Arroita et al., 2017). Alternatively, false-positive

28 occupancy models may be used to account for misclassification errors (Royle and Link,

29 2006; Chambert et al., 2015; Banner et al., 2018; Balantic and Donovan, 2019; Wright

30 et al., 2020). These models typically rely on combining the ambiguous ARU data with an

31 error-free source of unambiguous detections, or using a secondary method to validate de-

32 tections *post-hoc* (Chambert, Grant, Miller, Nichols, Mulder and Brand, 2018). However,

33 in each of these approaches, the probability a recording is assigned the correct or incorrect

34 species label is not explicitly modeled. As a result, erroneous detections are attributed to

35 an omnibus source, rather than to presence of a different species (as described in Wright

36 et al., 2020).

37   Additionally, both standard and false-positive occupancy models summarize counts of

38 detections as a binary response of detection or non-detection, thereby ignoring available in-

39 formation regarding relative activity or encounter frequencies. Estimates of relative activity

40 allow biologists to make inferences beyond occurrence, provide a more sensitive metric when

41 assessing population change, and better inform estimates of occurrence when false-positive

42 detections are present (Chambert, Waddle, Miller, Walls and Nichols, 2018; Wright et al.,

43 2020). Recently, false-positive occupancy models that simultaneously estimate misclassifi-

44 cation probabilities and relative activity have been developed (Kéry and Royle, 2021, ch. 7).

5

45   These so-called "coupled classification" models rely on unambiguous detections to inform

46   species-specific classification probabilities, which are used to adjust ambiguous detections

47   for false-positives (Wright et al., 2020; Spiers et al., 2021). By "coupling" ambiguous and

48   unambiguous detections from the same survey event within an integrated model, classifi-

49   cation rates are simultaneously estimated with occurrence and relative activity. However,

50   current ARU data post-processing workflows result in "uncoupled" auxiliary or calibration

51   data which are not co-located with the ambiguous detections. Consequently, classification

52   rates are estimated separately from occurrence and relative activity (Chambert et al., 2015;

53   Wright et al., 2020).

54   Due to multi-modality in the likelihood, all false-positive occupancy models require

55   additional information to identify detection parameters (Chambert et al., 2015). This

56   information can be incorporated in a variety of ways, including: 1) use of an error-free

57   validation method on a subset of recordings; 2) use of auxiliary or calibration data to es-

58   timate the multinomial classification probabilities; or 3) use of informative priors on the

59   classification probabilities in a Bayesian framework. In the former two cases, unambiguous

60   detection information is used to estimate the classification probabilities separately from

61   relative activity. These two approaches differ in that the error-free validation method is

62   used on *in situ* recordings from a subset of sites and visits, rather than relying on aux-

63   iliary or calibration data that are not necessarily representative of the same recording

64   conditions as the observed data. Calibration data are typically comprised of identified

65   acoustic recordings from known species, and therefore allow estimation of species-specific

66   classification probabilities. However, unlike the validated recordings, calibration data are

6

67 not collected simultaneously with the observed data. Therefore, we consider the validated

68 acoustic recordings as "coupled" with the ambiguous detection data, and the calibration

69 data as "uncoupled". Coupled information informs species occurrence and allows estima-

70 tion of classification probabilities assuming field conditions, rather than relying on voucher

71 recordings collected in non-characterized recording environments. However, the sample size

72 per species available to validate is constrained by the number of recordings identifiable to

73 species and the species-specific activity levels at a site. The factors most heavily influenc-

74 ing species-specific identification of recordings is site selection for ARU (e.g. open versus

75 forested), with microphone placement and orientation in relation to sources of ambient

76 noise contributing to the overall quality of the recordings which can compromise the iden-

77 tification process in software by obscuring the bat signal. Activity levels can fluctuate with

78 season, time of night, ambient conditions like temperature, moonlight, and precipitation,

79 and insect prey diversity and abundance.

80   Despite all false-positive models requiring additional sources of information to identify

81 model parameters, there is currently little to no guidance about how much information is

82 needed, or about how inferences may be impacted by using calibration versus validation

83 approaches to correct for the cross-species misclassifications. To investigate how unam-

84 biguous information impacts parameter estimation, we use unique empirical bioacoustic

85 data from British Columbia, Canada that surveyed 55 sites for 1-5 years and detected 17

86 species. These data are unique because all of the recordings that received a species label

87 from the automated classification process were subsequently reviewed by a human expert to

88 assign a species label or species group. Through these data, we explore how using coupled

89 validation data, using uncoupled calibration data, or relying on only informative priors af-

90 fects parameter identifiability and resulting uncertainty. Additionally, we explore common

91 choices of prior distributions within the Bayesian count detection framework and investi-

92 gate their impact on parameter estimation (bias, coverage, and posterior interval width).

93 Finally, we establish practical guidance on the quantity of unambiguous data required to

94 reduce post-processing costs without compromising parameter estimates.

2   Methods

2.1   Multi-species misclassification count detection model framework

98 Below, we briefly describe the count detection model framework; for greater detail, please

99 see Wright et al. (2020). The count detection model framework of Wright et al. (2020)

100 assumes the following notation. Let i index the site, let j index the visit to site i, and let

101 k index the species. The latent occupancy state of species k at site i, Z_{ik}, is represented

102 by a Bernoulli random variable:

$$Z_{ik} \sim \text{Bernoulli}(\psi_{ik}), \tag{1}$$

103 where ψ_{ik} represents the probability that species k occupies site i; occupancy states are

104 assumed independent across species. Site-specific covariates can be included through a

105 generalized linear model framework, $g(\psi_{ik}) = \boldsymbol{x}'_i \boldsymbol{\beta}_k$, where $g(.)$ represents an appropriate

106 link function, $\boldsymbol{x}'_i$ represents a row vector of site-level covariates, and $\boldsymbol{\beta}_k$ represents a vector

107 of regression coefficients for species k. Spatial or temporal dependence in occupancy among

108 species can be induced by incorporating hierarchical regression coefficients (Kéry and Royle,

109 2008; Spiers et al., 2021).

110   Conditional on site-level occupancy, the number of detections associated with species k

111 from site i on visit j, y_{ijk}, is modeled as a Poisson random variable:

$$Y_{ijk}|z_{ik} \sim \text{Poisson}(z_{ik}\lambda_{ijk}), \tag{2}$$

112 where λ_{ijk} represents the expected number of detections or encounter rate of species k

113 from visit j at site i. Visit-specific covariates can be included through a generalized linear

114 model framework, $g(\lambda_{ijk}) = \boldsymbol{v}'_{ij}\boldsymbol{\alpha}_k$, where $g(.)$ represents an appropriate link function, $\boldsymbol{v}'_{ij}$

115 represents a row vector of visit-specific covariates, and $\boldsymbol{\alpha}_k$ represents a vector of regression

116 coefficients for species k. Dependence in mean detection rates across species can again be

117 induced through hierarchical regression coefficients.

118   If the true species generating each detection is known without error, equations 1 and 2

119 fully describe the observed data. However, this is seldom the case as detections are often

120 incorrectly assigned a species label by automated software packages. To account for this

121 possibility, Wright et al. (2020) define a species confusion matrix, $\boldsymbol{\theta}$, in which element $\theta_{kk'}$

122 describes the probability that a detection truly belonging to species k is misidentified as

123 species k'. Then, due to properties of independent Poisson random variables, the total

124 number of recordings assigned a species k' label, $c_{ij.k'}$, is modeled as a Poisson random

9

125   variable:

$$c_{ij.k'} \sim \text{Poisson}\left(\sum_{k=1}^{K} z_{ik}\lambda_{ijk}\theta_{kk'}\right). \tag{3}$$

126   The multi-species count detection model reflects the observation process for acoustic data

127   by first modeling latent counts of audio recordings, then allocating those recordings to

128   observed species labels via the classification probabilities.

129   As with any false-positive model, the classification probability parameters are not iden-

130   tifiable from only ambiguous data and additional information is required to estimate model

131   parameters. If detections are validated after collection using an error-free method to assign

132   a species label to a recording, the count of detections truly belonging to species k that are

133   identified to species k' on visit j from site i are incorporated into the likelihood as Poisson

134   counts:

$$c_{ijkk'} \sim \text{Poisson}(z_{ik}\lambda_{ijk}\theta_{kk'}). \tag{4}$$

135   If the only available information about the software algorithms (in)accuracy for identifying

136   a species based on recording features is auxiliary or independent of the observed recordings,

137   the calibration data can be incorporated into the likelihood through a multinomial sampling

138   model (see Appendix A for model code).

139   **2.2   Acoustic bat monitoring in British Columbia, Canada**

140   Our work is motivated by a multi-species bat acoustic data set collected across 55 sites

141   in British Columbia, Canada between 2016 and 2020 (Figure 1); sites were monitored for

142   between one and five years. One to six stationary acoustic recording devices were deployed

10

143   within a 10 x 10 kilometer grid cell (a site) following the guidelines established for the

144   North American Bat Monitoring Program (NABat; Loeb et al., 2015). Within each site,

145   detectors were placed sufficiently far apart to minimize spatial dependence among recorded

146   calls from detectors within the same site (Loeb et al., 2015). Each recording device was

147   placed and activated for multiple consecutive nights and recorded echolocating bats between

148   sunset and sunrise nightly. Each detector was typically activated for seven nights, but some

149   detectors had as few as one or as many as 49 nights. To minimize the impact of runs in

150   bat activity and temporal dependence, only detections from the first and last night at each

151   detector were considered for analysis. While the count detection model can be adapted

152   to explicitly account for potential temporal dependence in detections, we simplified the

153   data structure to focus on the impacts of coupling unambiguous detection information on

154   parameter estimation.

155   Acoustic recordings were identified using the Kaleidoscope Pro acoustic classification

156   software for bats (`https://www.wildlifeacoustics.com`). All recordings with a Kaleidoscope-

157   assigned species label were then visually inspected (manually reviewed), and species labels

158   were either confirmed, changed to another species, or downgraded to a species group. The

159   analysis procedure followed the guidance of Reichert et al. (2018) such that auto-identified

160   labels were accepted only if the acoustic expert did not disagree with the identification.

161   In total, 17 bat species were identified. Following Wright et al. (2020), species that were

162   difficult to detect acoustically or that were not widespread, were combined into an "other"

163   category. This choice was made because we did not believe there was enough reliable

164   data to accurately estimate parameters associated with species in the "other" category.

165   Multi-species data sets are well-known for their over-abundance of zeros and a common

166   approach is to remove sparsely represented species within the assemblage. However, in

167   our case, the species within the "other" category can still be a source of false positives

168   and including them allows for a more complete representation of the observation process.

169   When constructing detection histories for each species, each night at a detector location

170   was considered an independent visit to the site.

171   These data are unique in that all acoustic recordings assigned an automated species

172   label received subsequent validation by a bat expert and, consequently, do not require false-

173   positive models for analyses. If one assumes that the manual validation process produces

174   close to 'error-free' identifications, then this data set provides a large number of recordings

175   to draw from that have an autoidentification species label and a subsequent 'error-free'

176   label. These data provide an opportunity to investigate the impact of validation effort on

177   parameter estimation by randomly masking the "error-free" validated species label for some

178   acoustic recordings. By doing so, we are able to generate typical acoustic data with varying

179   degrees of manual review, then fit models to each of these data sets in order to assess the

180   impact of the validation effort (Section 2.4). Additionally, a partially masked version of

181   the empirical data can be used to inform data generating values for a simulation study

182   designed to investigate the impact of coupling ambiguous and unambiguous detections on

183   model parameter estimates (Section 2.3).

184   For the occupancy portion of the model fit in the second simulation, the elevation (kilo-

185   meters), the annual mean temperature (degrees Celsius), and the annual mean precipitation

186   (millimeters) at the centroid of the 10×10 kilometer site were used; these covariates were

187 used to capture broad-scale heterogeneity in occupancy across the considered species. For

188 the relative activity portion of the model fit in the second simulation, the minimum nightly

189 temperature (degrees Celsius), the total precipitation (millimeters), and the percentage of

190 the moon illuminated by the sun (percent) at the centroid of the site for each visit were

191 used. All covariates were scaled to have a mean of zero and a standard deviation of one

192 prior to modeling. To account for potential changes in occupancy over time, the implicit

193 parameterization of dynamic occupancy models was used (Field et al., 2005; MacKenzie

194 et al., 2016). This parameterization accounts for temporal correlation in occupancy status

195 by incorporating year-specific intercepts. Both models were fit using NIMBLE (de Valpine

196 et al., 2017, 2020), with three independent chains of 5000 MCMC iterations each; both

197 models were assessed for convergence visually and through the Gelman-Rubin statistic
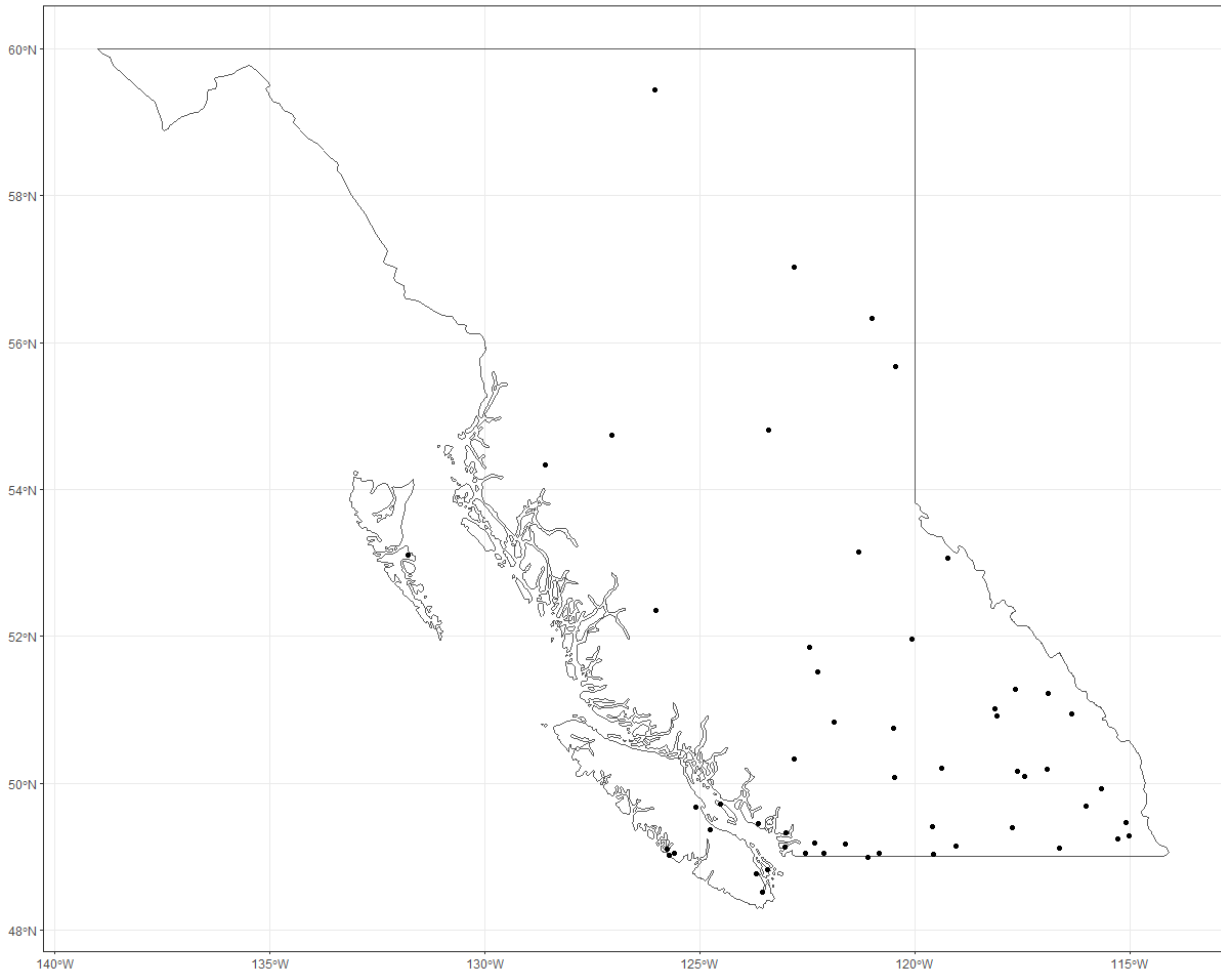
198 (Brooks and Gelman, 1998).

13

Figure 1: Monitoring locations in British Columbia, Canada. Each of the 55 10km × 10km NABat grid cells were surveyed with between one and six acoustic recording devices for between one and 49 nights for one to five years.

2.3 Simulation-based investigation into coupling and prior specification

The first simulation study investigated the impact of coupling ambiguous and unambiguous detections on Bayesian parameter estimation. Additionally, the first simulation explored how prior specification for the classification probabilities affects posterior distribution un-

14

$_{204}$ certainty and bias. We considered three potential formulations of the likelihood and up

$_{205}$ to three priors for each likelihood to account for multi-species misclassification (Table 1).

$_{206}$ The three likelihoods considered were: 1) the coupled count detection likelihood described

$_{207}$ in Section 2.1; 2) the uncoupled count detection likelihood that incorporates auxiliary

$_{208}$ or calibration data to estimate species classification probabilities; and 3) the uncoupled

$_{209}$ count detection likelihood that requires informative priors on classification probabilities.

$_{210}$ Each of these likelihoods were considered with vague priors for occupancy probability,

$_{211}$ $\psi_{ik} \sim \text{Beta}(1,1)$, and relative activity, $\lambda_{ijk} \sim N_{(0,\infty)}(0,100)$, and prior structures for the

$_{212}$ classification probabilities of the form $\boldsymbol{\theta}_{k,1:k} \sim \text{Dirichlet}(\boldsymbol{\alpha}^{(0)}_{k,1:k})$. The first two likelihoods

$_{213}$ were considered with three different prior specifications and the last likelihood was fit with

$_{214}$ a single informative prior to maintain parameter identifiability.

$_{215}$ The first prior considered places equal prior probability on each element of the classi-

$_{216}$ fication probability matrix, $\boldsymbol{\alpha}^{(0)} = \mathbf{1}$. This prior distribution, which can be thought of as

$_{217}$ a uniform distribution over the classification probability simplex, is commonly thought of

$_{218}$ as an uninformative prior distribution for multinomial probabilities, as it is analogous to

$_{219}$ a $\text{Beta}(1,1)$ prior distribution for Binomial probabilities. The second prior we consider is

$_{220}$ an informative Dirichlet prior that places a high degree of prior probability on the diag-

$_{221}$ onal elements of the confusion matrix; here, $\boldsymbol{\alpha}^{(0)} = \mathbf{1}$, with the diagonal elements of the

$_{222}$ matrix equal to 30. This prior reflects the assumption that the classification algorithm is

$_{223}$ more likely to correctly classify a recording, which is often the case with bat acoustic data.

$_{224}$ In general, the elements of $\boldsymbol{\alpha}^{(0)}$ can be interpreted as *a priori* classified counts; for each

$_{225}$ species, this prior translates to adding approximately 30 correctly classified detections and

$_{226}$ one incorrectly classified detection for each of the other species to the likelihood. Finally, we

$_{227}$ consider the reference distance prior for Dirichlet-multinomial sampling models proposed

$_{228}$ by Berger et al. (2015). This joint prior distribution is constructed to impose minimal prior

$_{229}$ information on the elements of the classification probability matrix. Here, $\boldsymbol{\alpha}^{(0)} = \frac{1}{k} \cdot \mathbf{1}$,

$_{230}$ where k is the number of species; see Berger et al. (2015) for more detail. The suite of

15

Impacts of coupling on modeling of acoustic data

231   models considered are summarized in Table 1. Only the informative prior was used in

232   conjunction with the uncoupled likelihood without calibration data, as the parameters are

233   not identifiable otherwise.

model	likelihood	prior
1	coupled	uniform on simplex
2	coupled	informative
3	coupled	reference distance
4	uncoupled with calibration data	uniform on simplex
5	uncoupled with calibration data	informative
6	uncoupled with calibration data	reference distance
7	uncoupled without calibration data	informative

Table 1: Description of models consider in simulation.

234   One hundred data sets were generated from the coupled version of the count detection

235   model as it best reflects the underlying ecological and observation processes for multispecies

236   acoustic data. Assumed parameter values were based on estimates from the bat acoustic

237   data collected annually in British Columbia, Canada between 2016 and 2020 (Appendix

238   B). Each generated data set had 100 sites with eight independent visits each. We assumed

239   two of the eight visits from each site were randomly selected for validation effort in which

240   all recordings with an assigned species label from the software were manually reviewed

241   by a bat expert. Calibration data sets for each iteration were created by separating the

242   validated detections from sample events and treating them as a uncoupled unambiguous

243   detections (see Appendix A for more detail).

244   All models were fit using the probabilistic programming language NIMBLE (de Valpine

245   et al., 2017, 2020); code is provided in Appendix A. Each model was run for 2500 MCMC

246   iterations, and the first 1250 iterations were discarded as warm-up. Each model was ran-

247   domly initialized near the data generating values in order to hasten convergence to the

248   posterior distribution. Initializing each model near the generating values ensured that all

249   models considered converged to the posterior distribution in a reasonable number of it-

250   erations; convergence of each model is required in order to make fair comparisons across

16

251 models. Even still, 32 of the 100 generated data sets resulted in the uncoupled model with

252 informative priors on the classification probabilities (Table 1, model 7) failing to converge;

253 these fitted models were excluded when summarizing the results. For each model, the

254 posterior mean parameter estimates were tracked, in addition to the associated 95% cred-

255 ibility interval and whether that interval captured the data generating values. See Section

256 2.2 for a full description of the data used to determine the simulation parameter values,

257 and Appendix B for simulation code.

2.4 Investigation into validation effort using empirical bat acoustic data

260 The second investigation explored the ramifications of validation effort on the coupled

261 count detection model parameter estimates using the empirical bat acoustic data set. To

262 explore how the number of sites and visits validated affects posterior estimates, we randomly

263 masked various proportions of manual labels from the acoustic data described in Section

264 2.2. This was done by first randomly selecting a subset of sites. Then, for those selected

265 sites, a subset of visits to retain the error-free manual species labels for all recordings was

266 randomly selected. All other visits had their manual species labels masked and the species

267 labels assigned by the software classifier were treated as ambiguous detections; the full suite

268 of scenarios considered is described in Table 2. The masking process was conducted three

269 different times for each scenario with different seeds for the randomization.

270   The hierarchical count detection model described in Section 2.2 with the reference

271 distance prior on classification probabilities was fit to each masked data set using NIMBLE

272 (de Valpine et al., 2017, 2020), including mean elevation (kilometers), annual precipitation

273 (millimeters), and annual mean temperature (degrees Celsius) as occupancy level predictors

274 and nightly temperature (degrees Celsius), nightly precipitation (millimeters), and percent

275 lunar illumination as relative activity level predictors. Each model was run for 5000 MCMC

17

276 iterations and the first 2500 iterations were discarded as warm-up. Posterior summaries

277 and convergence statistics were tracked for each model; models that failed to converge

278 were excluded. Following the simulation study, we present results from one example with

masking following the HM scenario (Table 2).

scenario	proportion sites conf.	proportion visits conf.	mean validated recordings
HH	1 (high)	0.75 (high)	53387
HM	1 (high)	0.5 (med)	34211
HL	1 (high)	0.25 (low)	19520
MH	0.75 (med)	0.75 (high)	42985
MM	0.75 (med)	0.5 (med)	24877
ML	0.75 (med)	0.25 (low)	14675
LH	0.5 (low)	0.75 (high)	28829
LM	0.5 (low)	0.5 (med)	18010
LL	0.5 (low)	0.25 (low)	10350

Table 2: Description of scenarios considered in simulation. The "proportion sites conf" column describes what proportion of sites received some form of manual vetting. The "proportion visits conf" column describes what proportion of visits were manually verified from the subset of sites that received manual review. The "mean validated recordings" column provides the mean total number of recordings that were validated across all species for the three simulated data sets for each scenario. Counts of manually validated calls by species for all scenarios is provided in Table 4 in Appendix B.

279

3　Results

3.1　Investigation into coupling and prior specification

282 Assuming 100 sites with 2 of 8 visits contributing "error-free" species detection counts, all

283 seven models resulted in posterior mean occupancy probability estimates that were unbiased

284 and produced 95% credibility intervals that achieved nominal coverage (Appendix B, Figure

285 10). In general, all models resulted in similar credibility interval width for occupancy

286 probability estimates, with the exception of the occupancy probability associated with the

287 "other" species category. For this species group, the coupled version of the model resulted

288 in the narrowest intervals, followed by the two uncoupled approaches (Appendix B, Figure

18

289   9).

290   Uncertainty in mean relative activity estimates differed across models. In Figure 2,

291   results from the simulation for a subset of species are provided; full results are provided

292   in Appendix B. In this section, we focus on four species that represent a spectrum of

293   occupancy (presence at a site) and activity (detections on a per night, per site basis);

294   1) western small-footed myotis (*Myotis ciliolabrum*, MYCI), 2) western long-eared myotis

295   (*Myotis evotis*, MYEV), 3) little brown myotis (*Myotis lucifugus*, MYLU), and 4) hoary bat

296   (*Lasiurus cinereus*, LACI). The generating values for occupancy probability and relative

297   activity rates associated with each of these groups ranged from the smallest to largest

      observed (Table 3).

species	occupancy probability	relative activity
LACI	0.61	4.16
MYCI	0.24	11.86
MYEV	0.70	2.41
MYLU	0.90	28.25

Table 3: Description of prevalence and relative activity of bat species considered. The occupancy probabilities and relative activity rates considered here range from the smallest to largest observed in the acoustic data from British Columbia, Canada.

298

299   Credibility interval widths for estimates from the count detection model with coupled

300   likelihood were narrower than for estimates from either model with uncoupled likelihoods,

301   regardless of the prior structure. On average, the model with coupled likelihood resulted in

302   credibility interval widths that were 18% narrower than the model with uncoupled auxiliary

303   data, and 36% narrower than the model with uncoupled likelihood without auxiliary data.

304   The model with uncoupled auxiliary data resulted in credibility interval widths that were,

305   on average, 21% narrower than the uncoupled model without auxiliary data. See Appendix

306   B for graphical summaries of credibility interval widths. On average, posterior mean rela-

307   tive activity estimates from both the coupled model and uncoupled model with auxiliary

308   data were unbiased and produced 95% credibility intervals that achieved nominal coverage,

309   regardless of the prior structure on the classification probabilities across all species. Rela-

310　tive activity estimates from the uncoupled model without auxiliary data resulted in biased

311　parameter estimates for some species and did not achieve nominal coverage rates.

312　　　　Credibility interval widths for classification probabilities were similar for the coupled

313　model and uncoupled model with auxiliary data. However, the uncoupled model without

314　calibration or auxiliary data (model 7) resulted in wider intervals (Figure 3). On average,

315　credibility intervals for the coupled model were 57% narrower than the uncoupled model

316　based on prior information; intervals for the uncoupled model based on auxiliary data were

317　56% narrower than the uncoupled model based on prior information, on average. The

318　coupled model and uncoupled model with auxiliary data resulted in the least bias and

319　greatest coverage for the uniform and informative priors. However, for both of these prior

320　structures, all models resulted in some bias and low coverage for diagonal elements of the

321　classification matrix, $\boldsymbol{\theta}$. Conversely, the reference distance prior resulted in the least bias

322　and highest coverage for all models considered.
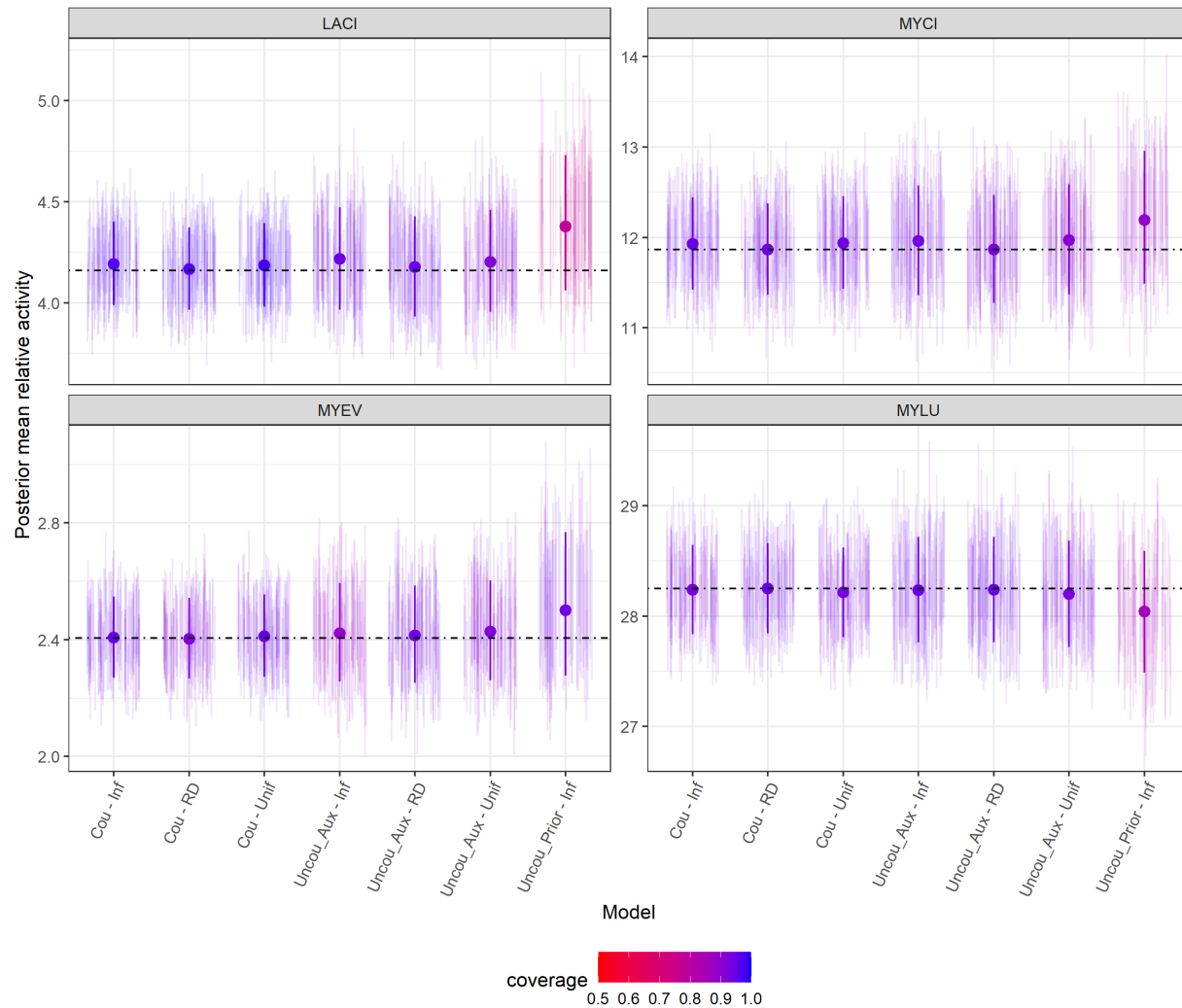
Figure 2: Ninety-five percent credibility intervals for relative activity parameter across all 100 simulations; average credibility intervals are displayed in bold. Color is determined by the proportion of simulated data sets for which credibility intervals captured the generating values. On average, the coupled versions of the model resulted in greater precision than did the uncoupled versions of the model. Additionally, only the reference distance prior resulted in unbiased parameter estimates and nominal coverage. The x-axis describes the combination of likelihood and prior; "cou" denotes the coupled version of the likelihood, "uncou_aux" denotes the uncoupled version of the likelihood with auxiliary data, "uncou_prior" denotes the uncoupled version of the likelihood without auxiliary data, "inf" refers to the informative prior, "RD" refers to the reference distance prior, and "unif" refers to the uniform prior.

Impacts of coupling on modeling of acoustic data
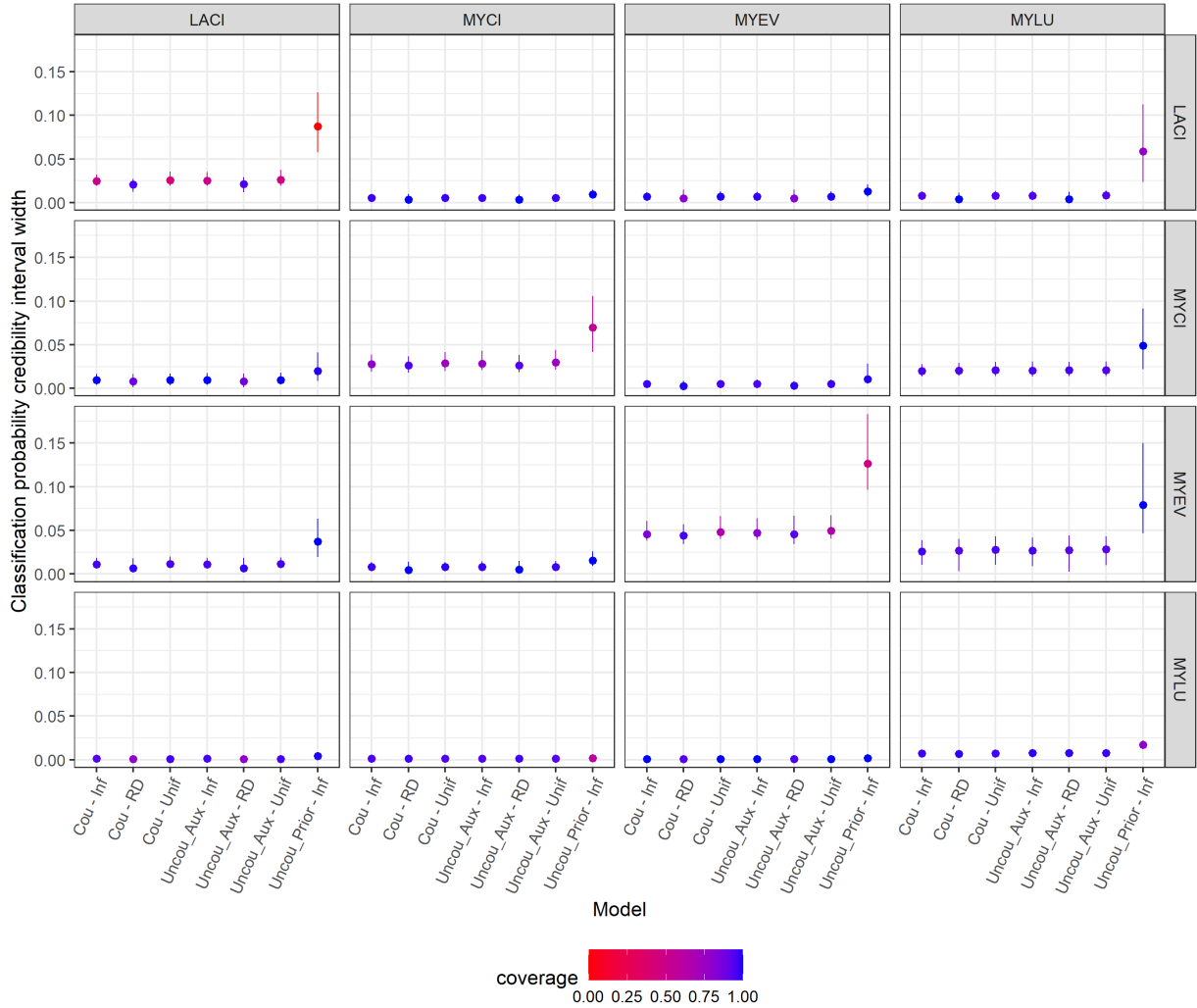
Figure 3: Credibility interval widths for classification probabilities across all 100 simulations; average credibility interval width is represented by the point. Color is determined by the proportion of simulated data sets for which credibility intervals captured the generating values. The coupled model and uncoupled model with auxiliary data resulted in the least uncertainty when estimating classification probabilities. Across all models including calibration data, on average, the reference distance prior structure resulted in the least bias and greatest coverage. The x-axis describes the combination of likelihood and prior; "cou" denotes the coupled version of the likelihood, "uncou_aux" denotes the uncoupled version of the likelihood with auxiliary data, "uncou_prior" denotes the uncoupled version of the likelihood without auxiliary data, "inf" refers to the informative prior, "RD" refers to the reference distance prior, and "unif" refers to the uniform prior.

22

3.2   Investigation into validation effort using empirical bat acoustic data

We present results for the four species described in Section 3.1, but provide all simulation results in Appendix B. We first note that the number of sites and visits with validated recordings affected model convergence. All fitted models for each of the four scenarios with the lowest validation effort failed to converge. Additionally, four other fitted models failed to converge: one from scenario MH, two from scenario MM, and one from scenario HL. In each of these cases, the lack of convergence was largely driven by classification probability estimates for the least active species (MYEV, MYVO, and the "other" category). Across the 16 fitted models that failed to converge, fewer than 800 recordings were selected for validation for the least active species; for nine of those fits, fewer than 400 recordings were looked at for validation for the least active species. In general, the lack of convergence was driven by an insufficient number of validated recordings for the less commonly detected species (Appendix B, Table 4). The fitted models that failed to converge are omitted from summary graphics.

Credibility interval width for occupancy-level coefficients were similar across all species and predictors, regardless of the quantity of error-free detections. In general, uncertainty was greater among coefficient estimates for less commonly detected species, but did not appear to vary with validation effort. Relative activity coefficient uncertainty also did not tend to vary with validation effort (Figure 4). For both the occupancy and relative activity coefficients, posterior intervals were generally similar in both center and width for all models that converged. Classification probability uncertainty varied with validation effort and species (Figure 5). In general, uncertainty was greatest among scenarios with lesser effort and among the least active species (MYEV). For the remaining species, each with greater activity, classification probability credibility interval width was similar, though marginally lesser for the two lowest effort scenarios that converged (MM, HL).

23

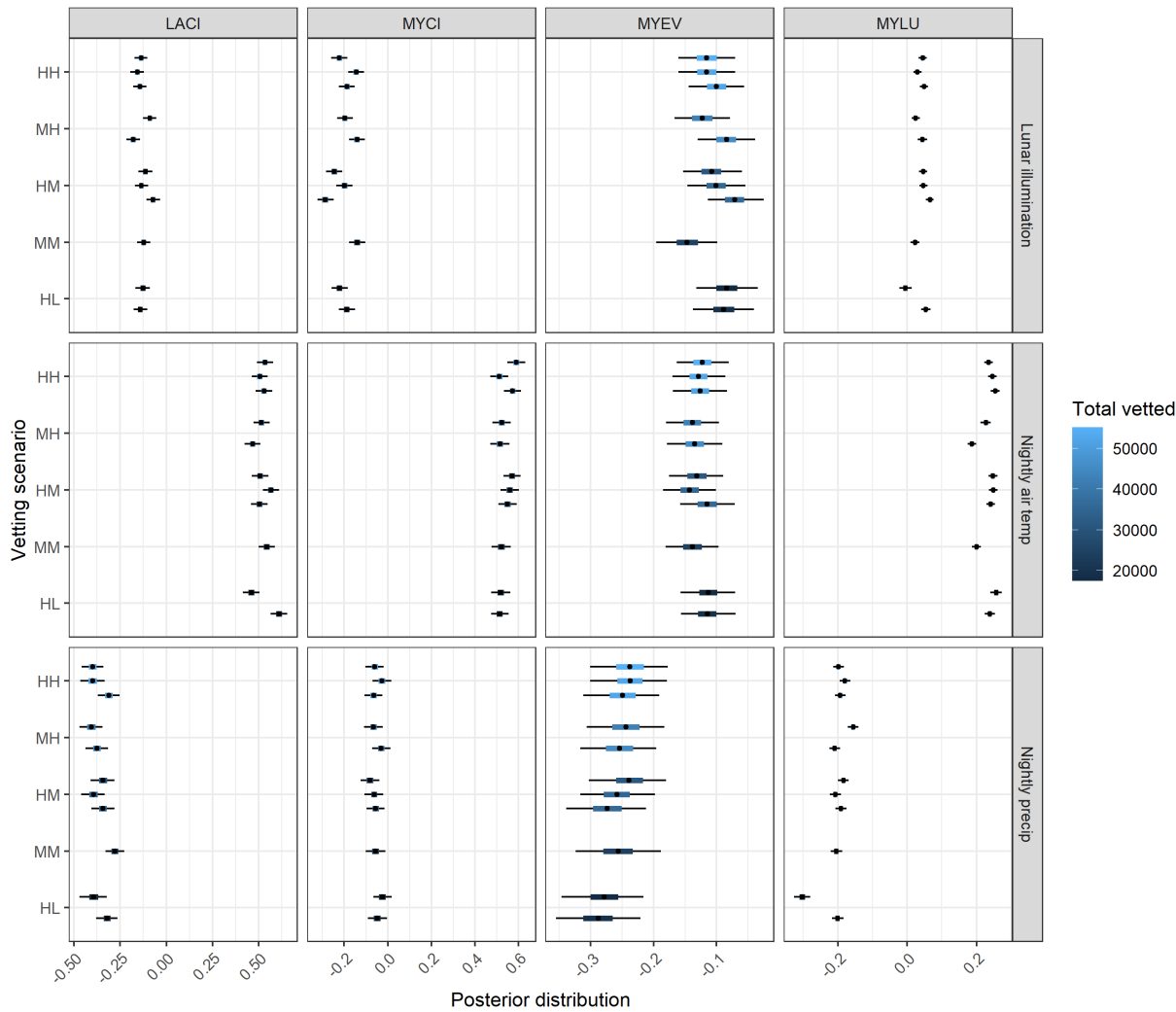*Impacts of coupling on modeling of acoustic data*

Figure 4: Posterior intervals for relative activity coefficients; points refer to the mean of the posterior distribution, thick lines to 50% credibility intervals, and thin lines to 95% credibility intervals. Validation scenarios are arranged from top to bottom by effort. Each scenario was run using three different seeds for randomization; missing intervals are indicative of lack of convergence. Additionally, all models fit to data from the ML, LH, LM, and LL scenarios failed to converge (omitted from plot).
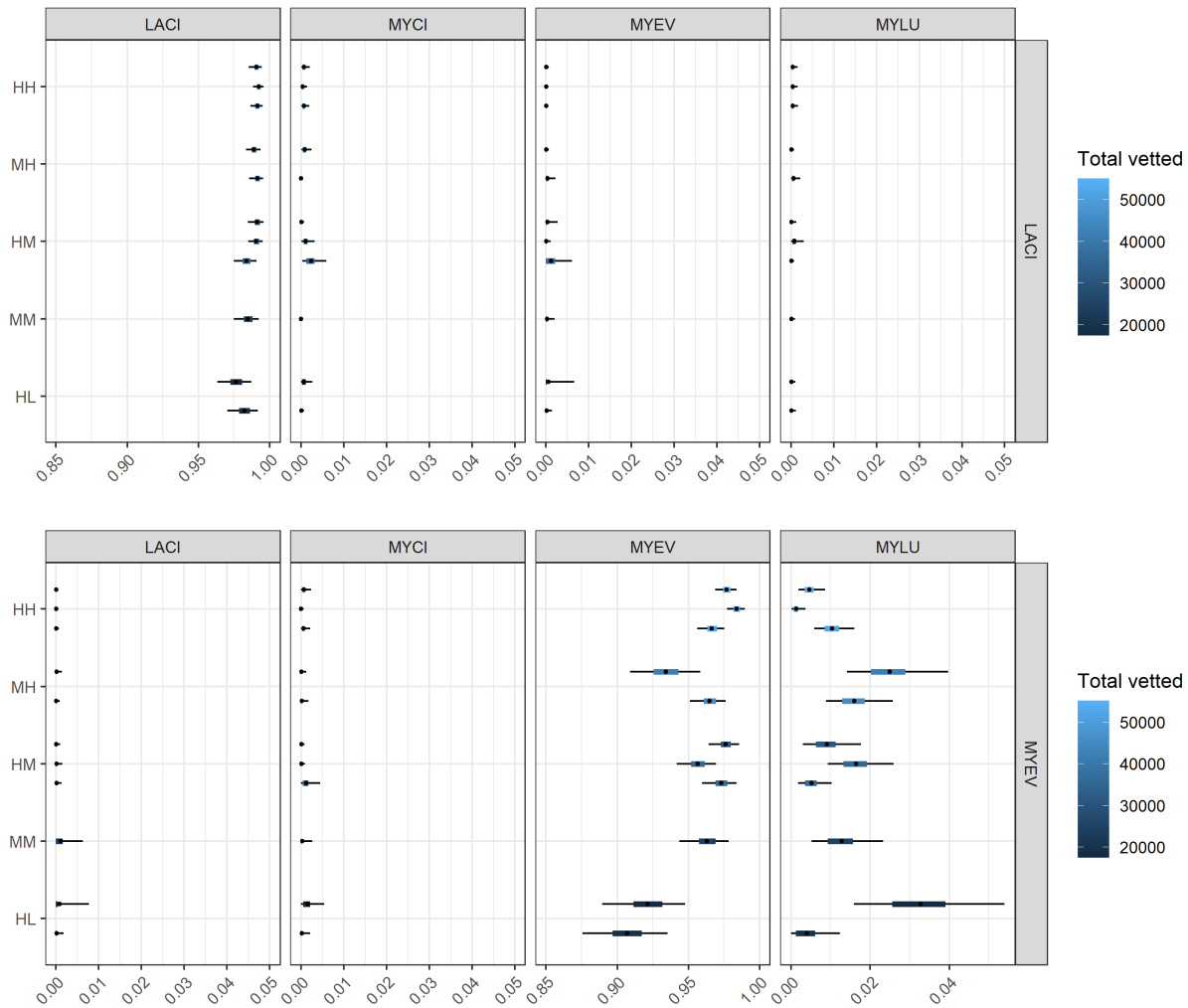
Figure 5: Posterior intervals for classification probabilities for species LACI and MYEV; points refer to the mean of the posterior distribution, thick lines to 50% credibility intervals, and thin lines to 95% credibility intervals. Validation scenarios are arranged from top to bottom by effort. Each scenario was run using three different seeds for randomization; missing intervals are indicative of lack of convergence. Additionally, all models fit to data from the ML, LH, LM, and LL scenarios failed to converge (omitted from plot). Rows indicate the true species and columns indicate the auto classified species.

3.3 Example results for scenario HM

In this section, we focus on one data set from the HM scenario, in which half of all re-visits from every site received complete manual validation, as this scenario resulted in the

352   lowest validation effort among the scenarios with consistent model convergence. Occu-

353   pancy probability coefficient estimates were similar across species for all four coefficients

354   associated with year, suggesting that most species behaved similarly over time (Figure 6).

355   Across all species, no meaningful changes in occupancy over time were detected. Inter-

356   cept estimates varied across species, in general reflecting naive occupancy estimates with

357   MYLU and LANO being most prevalent and MYCI being least prevalent. Across six of

358   the 10 species, occupancy was positively associated with annual mean temperature after

359   accounting for precipitation and elevation, suggesting increased prevalence with warmer

360   climates; the remaining four species shared a moderate negative association with tempera-

361   ture, though none of these credibility intervals included zero. Conversely, five of 10 species

362   shared a negative association with annual mean precipitation after accounting for temper-

363   ature and elevation, suggesting that occupancy probability decreases with wetter climates;

364   the remaining five species were moderately positively associated with precipitation, though

365   all credibility intervals included zero. Finally, adjusted for precipitation and temperature,

366   seven of the 10 species shared a positive association with elevation and credibility intervals

367   generally did not include zero for these species. Only MYVO, MYLU, and the "other"

368   species category shared a slight negative association with mean elevation, though these

369   credibility intervals did include zero.

370   Relative activity coefficient estimates varied across species for all covariates considered

371   (Figure 7), but most species shared the same directional associations. Intercept estimates

372   reflected naive activity rates, with MYLU and LANO being most active, and the conglom-

373   erate of less commonly detected species being least active. All species but MYCA shared

374   a negative association with nightly precipitation and all species but MYEV shared a pos-

375   itive association with nightly temperature, suggesting most species preferred warmer, dry

376   nights. Associations with lunar illumination varied across the species, with roughly half

377   of species sharing a positive association, the other half sharing a negative association, and

378   MYCA activity being unaffected by lunar illumination. However, only the other species

26

Impacts of coupling on modeling of acoustic data

group shared a strong association with lunar illumination, while the other species' associ-

ations were weaker. Across all species, correct classification probabilities were very high

(Figure 8), with all species exceeding 0.75. Only two classification errors were relatively

common: EPFU misclassified as LACI (posterior mean probability of 0.128) and MYVO

misclassified as MYLU (posterior probability of 0.135). The latter species reassignments

possibly stemming from features or artefacts in recordings that may be detected through

manual observation, but undetected through the auto-identification process.
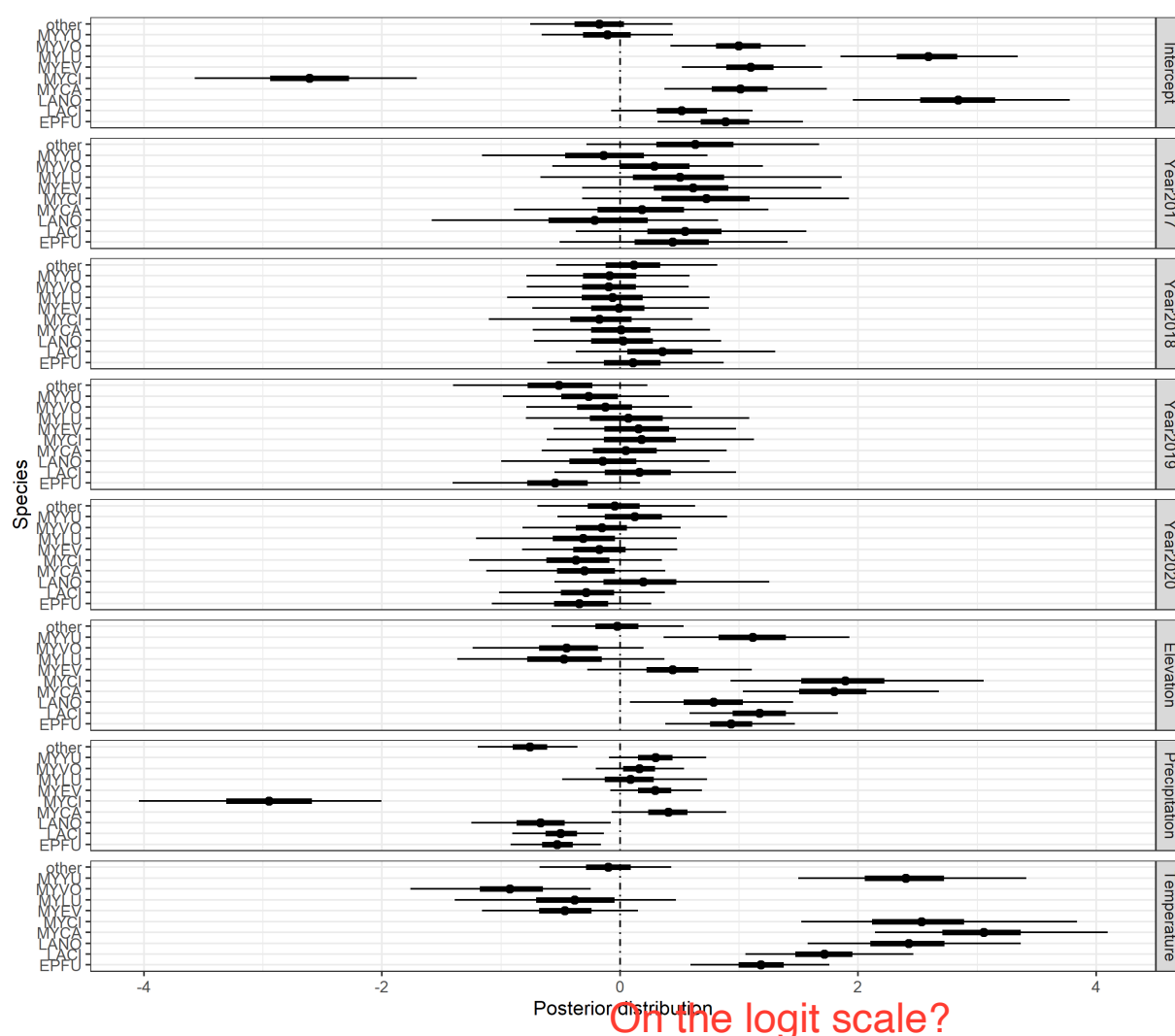
Figure 6: Posterior intervals for occupancy probability coefficient estimates; points refer to the mean of the posterior distribution, thick lines to 50% credibility intervals, and thin lines to 95% credibility intervals. The baseline intercept corresponds to 2016.

27

Impacts of coupling on modeling of acoustic data
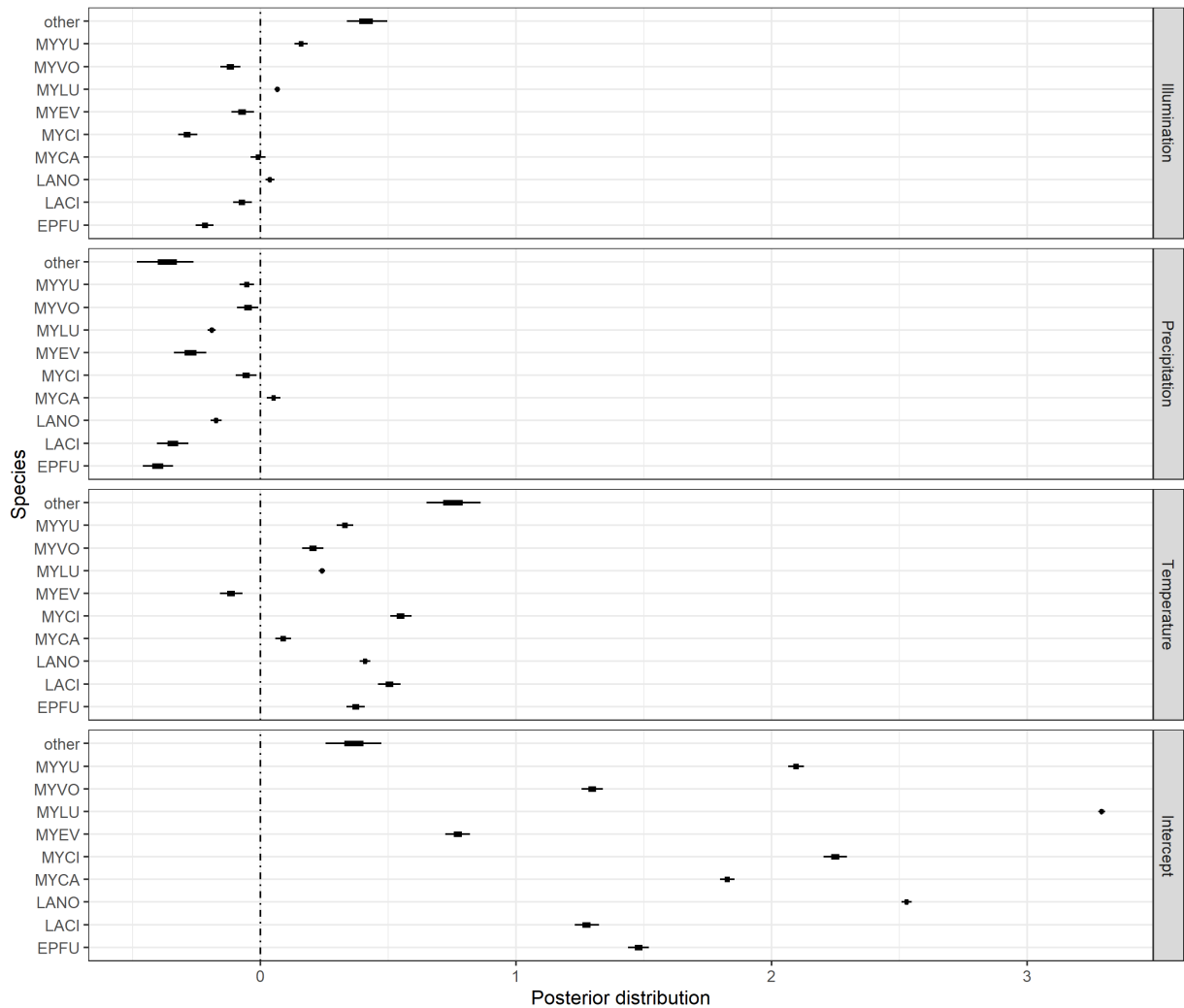
Figure 7: Posterior intervals for relative activity coefficient estimates; points refer to the mean of the posterior distribution, thick lines to 50% credibility intervals, and thin lines to 95% credibility intervals.
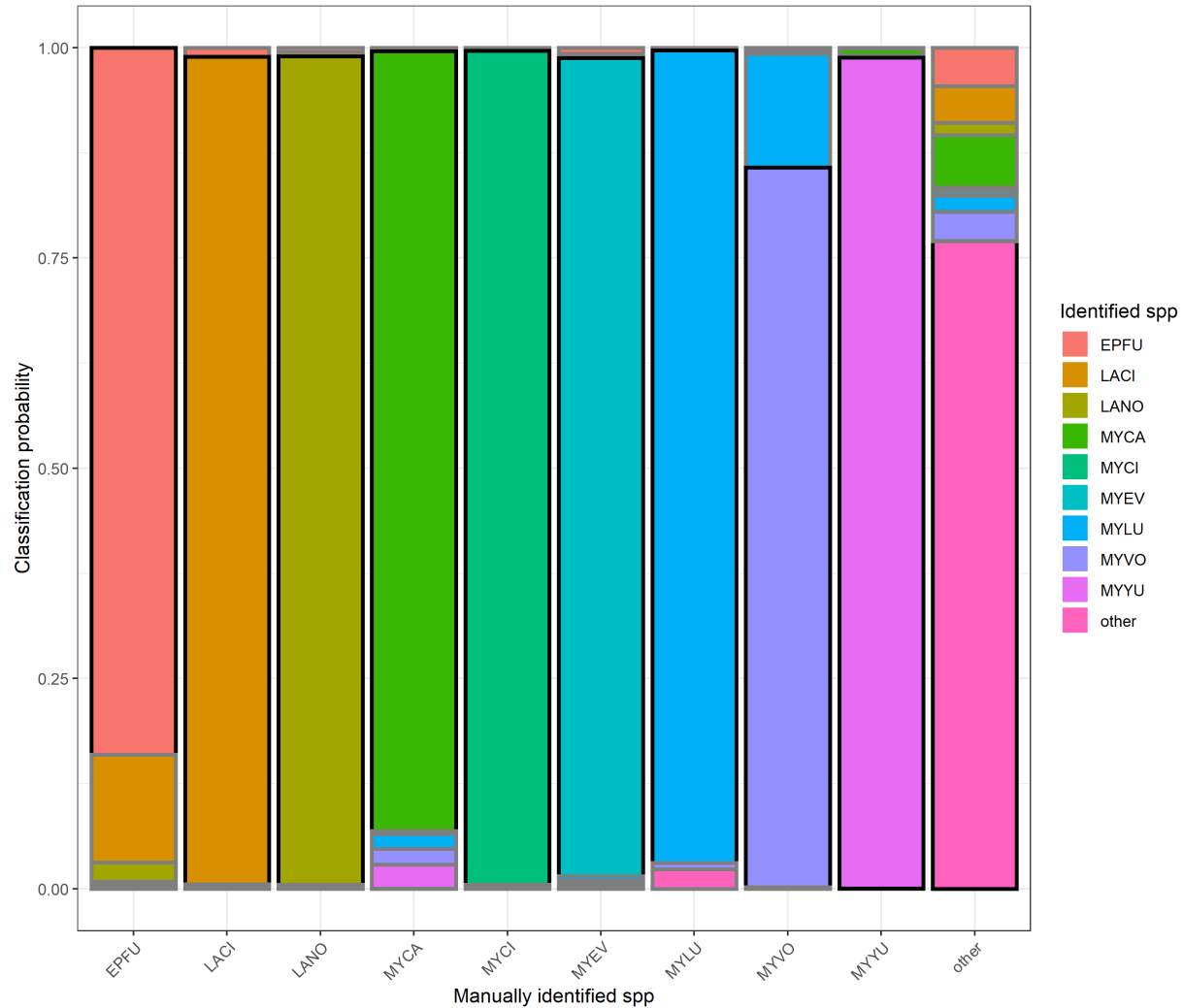
I think the confusion matrix would be clearer than this plot.

Figure 8: Posterior mean classification probabilities. The true call-generating species is on the x-axis and the identified species is determined by the fill color. "Correct" classification probabilities are outlined in black. High levels of agreement likely coincide with the analysis protocol followed (Reichert et al., 2018), and our exclusion of files that were manually labelled to be one of two or more potential species. For example, this would include all files labelled as EPFU or LANO and manually assigned to an EPFU/LANO label.

29

386 # 4   Discussion

387 Our simulation investigation demonstrated that simultaneous estimation of classification

388 probabilities and relative activity rates using the coupled count detection model results

389 in less bias and uncertainty when estimating model parameters, relative to approaches re-

390 lying on the informative Bayesian priors we considered or auxiliary data sets to estimate

391 classification probabilities. Additionally, we showed that count detection model parameter

392 estimates can be sensitive to prior specification on the classification probabilities, but that

393 the reference distance prior structure (Berger et al., 2015) results in unbiased parameter

394 estimates and nominal coverage rates for all model parameters. Our results indicated that

395 estimates of relative activity rates and classification probabilities benefited most from the

396 coupled model framework, while occupancy probability estimates were similar across all

397 modeling strategies. The similarity in occupancy probability estimates across modeling

398 strategies was likely related to the degree of concordance in assigned species labels between

399 the bat expert and software for the empirical data. Since empirical correct classification

400 probabilities for all species considered exceeded 0.80, few detections were required to con-

401 fidently confirm species presence at a site.

402   Within the coupled count detection model framework, we reviewed the impact of vari-

403 ous levels of detection validation effort on model parameter estimates. In general, we found

404 validation had little impact on parameter estimates, so long as there were enough validated

405 detections to identify model parameters. Lack of identifiability in false-positive occupancy

406 model parameters is a known issue, and sufficient information about classification probabil-

407 ities is required to resolve the problem (Chambert et al., 2015). For some of the scenarios

408 we considered in our simulation, there were not enough validated detections per species,

409 leading to convergence issues due to lack of parameter identifiability (see Table 4 in Ap-

410 pendix B for details). For the empirical bat data, manually validating half of revisits from

411 every site resulted in enough unambiguous information to identify model parameters.

412   The convergence issues experienced in the second simulation investigating validation

413   effort were largely driven by the least common species. In that simulation, sites and

414   visits were selected for complete manual validation randomly. As a result, less commonly

415   recorded species were less likely to be included in the manually validated visits as the

416   validation effort decreased, leading to an insufficient number of validated recordings for

417   those species. In practice, a strategic manual validation process, in which a certain quota

418   of recordings is randomly selected and validated for each species, can help reduce validation

419   effort without compromising parameter identifiability. As a consequence of the quota-

420   sampling process, only a subset of recordings from some visits would receive an error-free

421   species label assignment. The coupled count-detection model as described in Section 2.1

422   requires manual validation of all detections at a visit level, as responses are summarized

423   as counts of detections per visit. However, an unaggregated version of of this model could

424   be used to model individual detections, and therefore accommodate partial validation of

425   recordings for a subset of visits. This remains an area of active research. In the future, we

426   plan to investigate the impact of this quota validation system on parameter estimates and

427   investigate how the required quota may be influenced by occupancy probabilities, relative

428   activity, and classification probabilities.

429   In practice, manual validation of automated species identifications accounts for a sig-

430   nificant portion of the cost associated with acoustic monitoring programs. As a result,

431   validation effort may be dictated by budget constraints. Despite this, given fixed valida-

432   tion effort, we have shown that coupling unambiguous and ambiguous detections results in

433   less bias and uncertainty in estimates of relative activity and species classification probabil-

434   ities. However, monitoring programs should consider their measurable objectives of interest

435   when balancing cost of validation versus reducing uncertainty in parameter estimates. For

436   example, when estimating occupancy probabilities, the count detection model performed

437   similarly when using both the coupled and uncoupled likelihood structures. Additionally,

438   if the core question of interest is to investigate patterns in occupancy over space and time

439 but the species-specific patterns in relative activity are not a primary concern, there ex-

440 ist other false-positive occupancy models that require less rigorous validation (Chambert

441 et al., 2015). When designing survey and validation effort, these factors should be consid-

442 ered collectively.

443   Long-term monitoring programs may also find it challenging to maintain the level of

444 validation effort required to fit the coupled model annually. In these situations, validation

445 effort can be reduced by instead creating a comprehensive calibration data set to be used

446 in the uncoupled framework. When developing a calibration data set, care should be taken

447 to try and replicate *in situ* recording conditions, as the calibration data set is assumed

448 representative of the classification process for all surveyed sites. For example, for bat

449 acoustic surveys, exploring whether (mis)classification probabilities vary by local recording

450 habitat (e.g., forest interior versus flyways) or microphone deployment. Additionally, our

451 simulations investigating validation effort assumed the least informative reference distance

452 prior structure on classification probabilities. While the reference prior results in unbiased

453 parameter estimates, empirically informed priors could be used for rarer species if there

454 is insufficient validation effort available; when possible, these subjective priors should be

455 based on existing knowledge about the software algorithm's accuracy for a specific study

456 region. Finally, rare and/or difficult-to-record species can be aggregated into a single

457 species category when fitting any version of the count-detection model. Aggregating these

458 species can increase the minimum number of validated detections across all species, thereby

459 improving parameter identifiability.

460   Acoustic recording units continue to gain traction as a cost-effective source of multi-

461 species, range-wide monitoring data for multiple taxonomic groups. Analyses of ARU-based

462 survey data requires statistical techniques that accommodate the uncertainty in species la-

463 bels that is inherent to the automated classification process. Our results suggest that, when

464 possible, acoustic surveys should rely on coupled validated detection information to account

465 for false-positive detections, rather than uncoupled calibration data sets. However, if the

466 assemblage of interest contains a large number of rarely detected and less prevalent species,

467 an intractable amount of effort may be required, suggesting there are benefits to curating a

468 calibration data set that is representative of the observation process. Our findings provide

469 insights into the practical challenges associated with statistical analyses of bat activity and

470 occupancy using ARU data, and possible analytical solutions to support reliable and cost

471 effective monitoring of wildlife in the face of known sources of observation errors.

Impacts of coupling on modeling of acoustic data

Author contributions

CS and KI conceived the ideas and designed the simulations; CL and JR collected and collated the data; CS led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

Acknowledgments

Data accessibility

All data, simulation code, and model code used in this manuscript are archived on GitHub at `https://github.com/StrattonCh/CoupledUncoupled`.

References

Balantic, C. and Donovan, T. (2019), 'Dynamic wildlife occupancy models using automated acoustic monitoring data', *Ecological Applications* **29**(3), e01854.

URL: *https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1002/eap.1854*

Banner, K. M., Irvine, K. M., Rodhouse, T. J., Wright, W. J., Rodriguez, R. M. and Litt, A. R. (2018), 'Improving geographically extensive acoustic survey designs for modeling species occurrence with imperfect detection and misidentification', *Ecology and Evolution* **8**(12), 6144–6156.

URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/ece3.4162*

Baumgardt, J., Morrison, M., Brennan, L., Pierce, B. and Campbell, T. (2018), 'Development of multispecies, long-term monitoring programs for resource management', *Rangeland Ecology & Management* **72**.

Beason, R., Riesch, R. and Koricheva, J. (2019), 'AURITA: an affordable, autonomous recording device for acoustic monitoring of audible and ultrasonic frequencies', *Bioacoustics* **28**, 381–396.

Berger, J. O., Bernardo, J. M. and Sun, D. (2015), 'Overall objective priors', *Bayesian Analysis* **10**(1).

URL: *http://dx.doi.org/10.1214/14-BA915*

Brooks, S. and Gelman, A. (1998), 'General methods for monitoring convergence of iterative simulations', *Journal of Computational and Graphical Statistics* **7**, 434–455.

Chambert, T., Grant, E. H. C., Miller, D. A. W., Nichols, J. D., Mulder, K. P. and Brand, A. B. (2018), 'Two-species occupancy modelling accounting for species misidentification and non-detection', *Methods in Ecology and Evolution* **9**(6), 1468–1477.

URL: *https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12985*

35

Chambert, T., Miller, D. A. W. and Nichols, J. D. (2015), 'Modeling false-positive detections in species occurrence data under different study designs', *Ecology* **96**(2), 332–339.

URL: *https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1890/14-1507.1*

Chambert, T., Waddle, J. H., Miller, D. A. W., Walls, S. C. and Nichols, J. D. (2018), 'A new framework for analysing automated acoustic species detection data: Occupancy estimation and optimization of recordings post-processing', *Methods in Ecology and Evolution* **9**(3), 560–570.

URL: *https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12910*

de Valpine, P., Paciorek, C., Turek, D., Michaud, N., Anderson-Bergman, C., Obermeyer, F., Wehrhahn Cortes, C., Rodrìguez, A., Temple Lang, D. and Paganin, S. (2020), *NIMBLE: MCMC, Particle Filtering, and Programmable Hierarchical Modeling*. R package version 0.10.1.

URL: *https://cran.r-project.org/package=nimble*

de Valpine, P., Turek, D., Paciorek, C., Anderson-Bergman, C., Temple Lang, D. and Bodik, R. (2017), 'Programming with models: writing statistical algorithms for general model structures with NIMBLE', *Journal of Computational and Graphical Statistics* **26**, 403–413.

Field, S. A., Tyre, A. J. and Possingham, H. P. (2005), 'Optimizing allocation of monitoring effort under economic and observational constraints', *Journal of Wildlife Management* **69**(2), 473–482.

URL: *https://doi.org/10.2193/0022-541X(2005)069[0473:OAOMEU]2.0.CO;2*

Gibb, R., Browning, E., Glover-Kapfer, P. and Jones, K. E. (2019), 'Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring', *Methods in Ecology and Evolution* **10**(2), 169–185.

URL: *https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13101*

541 Guillera-Arroita, G., Lahoz-Monfort, J. J., van Rooyen, A. R., Weeks, A. R. and Tingley,
542   R. (2017), 'Dealing with false-positive and false-negative errors about species occurrence
543   at multiple levels', *Methods in Ecology and Evolution* **8**(9), 1081–1091.
544   **URL:** *https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12743*

545 Kéry, M. and Royle, J. A. (2008), 'Hierarchical Bayes estimation of species richness and
546   occupancy in spatially replicated surveys', *Journal of Applied Ecology* **45**(2), 589–598.

547 Kéry, M. and Royle, J. A. (2021), *Applied hierarchical modeling in ecology: analysis of
548   distribution, abundance and species richness in R and BUGS*, Vol. 2, Academic Press.

549 Loeb, S., Rodhouse, T., Ellison, L., Lausen, C., Reichard, J., Irvine, K., Ingersoll, T., Cole-
550   man, J., Thogmartin, W., Sauer, J., Francis, C., Bayless, M., Stanley, T. and Johnson,
551   D. (2015), 'A plan for the North American Bat Monitoring Program (NABat)'.

552 MacKenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Andrew Royle, J. and
553   Langtimm, C. A. (2002), 'Estimating site occupancy rates when detection probabilities
554   are less than one', *Ecology* **83**(8), 2248–2255.

555 MacKenzie, D. I., Nichols, J. D., Royle, A. J., Pollock, K. H., Bailey, L. L. and Hines,
556   J. E. (2016), *Occupancy estimation and modeling: inferring patterns and dynamics of*,
557   Elsevier Academic Press.

558 Measey, G. J., Stevenson, B. C., Scott, T., Altwegg, R. and Borchers, D. L. (2017), 'Count-
559   ing chirps: acoustic monitoring of cryptic frogs', *Journal of Applied Ecology* **54**(3), 894–
560   902.
561   **URL:** *https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/1365-2664.12810*

562 Morganti, M., Manica, M., Bogliani, G., Gustin, M., Luoni, F., Trotti, P., Perin, V. and
563   Brambilla, M. (2019), 'Multi-species habitat models highlight the key importance of

564 flooded reedbeds for inland wetland birds: Implications for management and conserva-
565 tion', *Avian Research* **10**, 15.

566 Newson, S. E., Bas, Y., Murray, A. and Gillings, S. (2017), 'Potential for coupling the
567 monitoring of bush-crickets with established large-scale acoustic monitoring of bats',
568 *Methods in Ecology and Evolution* **8**(9), 1051–1062.
569 **URL:** *https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12720*

570 Reichert, B., Bayless, M., Cheng, T., Coleman, J., Francis, C., Frick, W., Gotthold, B.,
571 Irvine, K., Lausen, C., Li, H., Loeb, S., Reichard, J., Rodhouse, T., Segers, J., Siemers,
572 J., Thogmartin, W. and Weller, T. (2021), 'NABat: a top-down, bottom-up solution to
573 collaborative continental-scale monitoring', *Ambio* **50**.

574 Reichert, B., Lausen, C., Loeb, S., Weller, T., Allen, R., Britzke, E., Hohoff, T., Siemers,
575 J., Burkholder, B., Herzog, C. and Verant, M. (2018), 'A guide to processing bat acoustic
576 data for the North American Bat Monitoring Program (NABat)', *U.S. Geological Survey*
577 *Open-File Report* pp. 1–43.

578 Royle, J. A. and Link, W. A. (2006), 'Generalized site occupancy models allowing for false
579 positive and false negative errors', *Ecology* **87**(4), 835–841.

580 Shonfield, J. and Bayne, E. (2017), 'Autonomous recording units in avian ecological re-
581 search: Current use and future applications', *Avian Conservation and Ecology* **12**, 14.

582 Spiers, A., Royle, J., Torrens, C. and Joseph, M. (2021), 'Estimating occupancy dynamics
583 and encounter rates with species misclassification: a semi-supervised individual-level
584 approach'. BioRxiv preprint.

585 Wright, W. J., Irvine, K. M., Almberg, E. S. and Litt, A. R. (2020), 'Modelling misclassi-
586 fication in multi-species acoustic data when estimating occupancy and relative activity',

587　　*Methods in Ecology and Evolution* **11**(1), 71–81.

588　　**URL:** *https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13315*

Appendix A: Count detection model code

Below, we provide NIMBLE code for each of the models fit in the manuscript. Complete code, including code to simulate from each model and the data arguments for each model, are provided digitally with the simulation scripts at `https://github.com/StrattonCh/CoupledUncoupled`.

Coupled count detection model

```
# priors
for(k in 1:nspecies){
  psi[k] ~ dbeta(1, 1)
  lambda[k] ~ T(dnorm(0, sd = 100), 0, Inf)
  theta[k, 1:nspecies] ~ ddirch(alpha = alpha0[k, 1:nspecies])
}

# likelihood - site level occupancy
for(site in 1:nsites){
  for(k in 1:nspecies){
    z[site,k] ~ dbern(psi[k])
  }
}

# likelihood 1 - unambiguous detections
for(row in 1:n1){
  y1[row] ~ dpois(
    z[site1[row],true_spp1[row]] *
    lambda[true_spp1[row]] *
    theta[true_spp1[row],id_spp1[row]]
  )
}

# likelihood 2 - ambiguous detections
for(row in 1:n2){
  for(true in 1:nspecies){
    mean2[row, true] <- z[site2[row],true]*lambda[true]*theta[true,id_spp2[row]]
  }
  y2[row] ~ dpois(sum(mean2[row, 1:nspecies]))
}
```

625 ## Uncoupled count detection model based on auxiliary data

```
626  # priors
627  for(k in 1:nspecies){
628    psi[k] ~ dbeta(1, 1)
629    lambda[k] ~ T(dnorm(0, sd = 100), 0, Inf)
630    theta[k, 1:nspecies] ~ ddirch(alpha = alpha0[k, 1:nspecies])
631
632    # aux counts to inform theta
633    aux_counts[k, 1:nspecies] ~ dmulti(size = aux_size[k], prob = theta[k, 1:nspecies])
634  }
635
636  # likelihood - site level occupancy
637  for(site in 1:nsites){
638    for(k in 1:nspecies){
639      z[site,k] ~ dbern(psi[k])
640    }
641  }
642
643  # likelihood 2 - ambiguous detections
644  for(row in 1:n2){
645    for(true in 1:nspecies){
646      mean2[row, true] <- z[site2[row],true]*lambda[true]*theta[true,id_spp2[row]]
647    }
648    y2[row] ~ dpois(sum(mean2[row, 1:nspecies]))
649  }
```

650 ## Uncoupled count detection model based on priors

```
651  # priors
652  for(k in 1:nspecies){
653    psi[k] ~ dbeta(1, 1)
654    lambda[k] ~ T(dnorm(0, sd = 100), 0, Inf)
655    theta[k, 1:nspecies] ~ ddirch(alpha = alpha0[k, 1:nspecies])
656  }
657
658  # likelihood - site level occupancy
659  for(site in 1:nsites){
660    for(k in 1:nspecies){
661      z[site,k] ~ dbern(psi[k])
662    }
663  }
```

41

```
664
665  # likelihood 2 - ambiguous detections
666  for(row in 1:n2){
667    for(true in 1:nspecies){
668      mean2[row, true] <- z[site2[row],true]*lambda[true]*theta[true,id_spp2[row]]
669    }
670    y2[row] ~ dpois(sum(mean2[row, 1:nspecies]))
671  }
```

Hierarchical count detection model

```
673  # priors
674  for(p in 1:p_beta){
675    mu_beta[p] ~ dnorm(0, sd = 2)
676    sigma2_beta[p] ~ T(dnorm(0, sd = 2), 0, Inf)
677  }
678
679  for(p in 1:p_beta){
680    for(k in 1:nspecies){
681      beta[p, k] ~ dnorm(mu_beta[p], var = sigma2_beta[p])
682    }
683  }
684
685  for(p in 1:p_alpha){
686    mu_alpha[p] ~ dnorm(0, sd = 2)
687    sigma2_alpha[p] ~ T(dnorm(0, sd = 2), 0, Inf)
688  }
689
690  for(p in 1:p_alpha){
691    for(k in 1:nspecies){
692      alpha[p, k] ~ dnorm(mu_alpha[p], var = sigma2_alpha[p])
693    }
694  }
695
696  for(k in 1:nspecies){
697    theta[k, 1:nspecies] ~ ddirch(alpha = alpha0[k, 1:nspecies])
698  }
699
700  # likelihood - site level occupancy
701  for(site in 1:nsites){
702    for(k in 1:nspecies){
703      logit(psi[site, k]) <- (beta[1:p_beta, k] %*% X[site, 1:p_beta])[1,1]
704      z[site, k] ~ dbern(psi[site, k])
```

```
705      }
706    }
707
708    # likelihood 1 - unambiguous detections
709    for(row in 1:n1){
710      # calculate lambda
711      log(lambda[row]) <- (alpha[1:p_alpha, true_spp1[row]] %*% W1[row, 1:p_alpha])[1,1]
712
713      y1[row] ~ dpois(
714        z[site1[row],true_spp1[row]] *
715        lambda[row] *
716        theta[true_spp1[row],id_spp1[row]]
717      )
718    }
719
720    # likelihood 2 - ambiguous detections
721    for(row in 1:n2){
722      for(true in 1:nspecies){
723        # calculate lambda
724        log(lambda2[row, true]) <- (alpha[1:p_alpha, true] %*% W2[row, 1:p_alpha])[1,1]
725        mean2[row, true] <- z[site2[row],true]*lambda2[row, true]*theta[true,id_spp2[row]]
726      }
727      y2[row] ~ dpois(sum(mean2[row, 1:nspecies]))
728    }
```

43

Appendix B: Simulation details

Below, we provide additional figures and simulation details that were omitted from the main text. All code used to conduct these simulations and generate the accompanying figures are provided digitally with the simulation scripts at `https://github.com/StrattonCh/CoupledUncoupled`.

Data generating values

Simulation parameters are based on the posterior mean estimates from the count detection model fit to acoustic data from British Columbia, Canada.

$$
\psi = \begin{matrix} \text{EPFU} \\ \text{LACI} \\ \text{LANO} \\ \text{MYCA} \\ \text{MYCI} \\ \text{MYEV} \\ \text{MYLU} \\ \text{MYVO} \\ \text{MYYU} \\ \text{other} \end{matrix} \begin{bmatrix} 0.6331 \\ 0.6122 \\ 0.8490 \\ 0.6972 \\ 0.2365 \\ 0.7036 \\ 0.8978 \\ 0.7018 \\ 0.5030 \\ 0.4160 \end{bmatrix}, \qquad \lambda = \begin{matrix} \text{EPFU} \\ \text{LACI} \\ \text{LANO} \\ \text{MYCA} \\ \text{MYCI} \\ \text{MYEV} \\ \text{MYLU} \\ \text{MYVO} \\ \text{MYYU} \\ \text{other} \end{matrix} \begin{bmatrix} 5.9347 \\ 4.1603 \\ 14.2532 \\ 6.1985 \\ 11.8649 \\ 2.4050 \\ 28.2490 \\ 3.8890 \\ 8.9623 \\ 1.9783 \end{bmatrix}
$$

44

Impacts of coupling on modeling of acoustic data

$$
\Theta =
\begin{array}{c}
 \\
\text{EPFU} \\
\text{LACI} \\
\text{LANO} \\
\text{MYCA} \\
\text{MYCI} \\
\text{MYEV} \\
\text{MYLU} \\
\text{MYVO} \\
\text{MYYU} \\
\text{other}
\end{array}
\begin{array}{|cccccccccc|}
\text{EPFU} & \text{LACI} & \text{LANO} & \text{MYCA} & \text{MYCI} & \text{MYEV} & \text{MYLU} & \text{MYVO} & \text{MYYU} & \text{other} \\
0.7878 & 0.1764 & 0.0245 & 0.0011 & 0.0005 & 0.0032 & 0.0014 & 0.0004 & 0.0006 & 0.0040 \\
0.0090 & 0.9826 & 0.0018 & 0.0016 & 0.0007 & 0.0017 & 0.0009 & 0.0006 & 0.0007 & 0.0005 \\
0.0071 & 0.0051 & 0.9829 & 0.0004 & 0.0002 & 0.0002 & 0.0028 & 0.0001 & 0.0004 & 0.0006 \\
0.0007 & 0.0005 & 0.0029 & 0.9235 & 0.0024 & 0.0008 & 0.0224 & 0.0183 & 0.0280 & 0.0005 \\
0.0026 & 0.0037 & 0.0005 & 0.0018 & 0.9691 & 0.0006 & 0.0185 & 0.0009 & 0.0010 & 0.0012 \\
0.0039 & 0.0014 & 0.0022 & 0.0014 & 0.0009 & 0.9493 & 0.0181 & 0.0091 & 0.0035 & 0.0101 \\
0.0002 & 0.0003 & 0.0002 & 0.0009 & 0.0015 & 0.0005 & 0.9650 & 0.0071 & 0.0003 & 0.0241 \\
0.0011 & 0.0008 & 0.0031 & 0.0021 & 0.0031 & 0.0014 & 0.1531 & 0.8327 & 0.0014 & 0.0011 \\
0.0009 & 0.0006 & 0.0008 & 0.0119 & 0.0003 & 0.0007 & 0.0043 & 0.0006 & 0.9789 & 0.0009 \\
0.0560 & 0.0389 & 0.0055 & 0.0097 & 0.0045 & 0.0128 & 0.0348 & 0.0077 & 0.0018 & 0.8285
\end{array}
$$

Impacts of coupling on modeling of acoustic data

scenario	seed	EPFU	LACI	LANO	MYCA	MYCI	MYEV	MYLU	MYVO	MYYU	other
HH	1000	2737	2917	10933	3471	2574	1492	20753	1921	4188	1305
HH	2000	2525	3161	10593	3715	2369	1412	21719	2218	3971	1118
HH	3000	2781	3176	10435	3982	2592	1531	22283	2131	4847	1310
HM	1000	1648	1686	7384	2471	1556	871	14791	1560	3267	899
HM	2000	1772	2428	6740	2265	1762	1065	13815	1665	3017	797
HM	3000	1591	1643	5356	2849	927	975	12421	1432	3260	721
HL	1000	1140	1664	4606	1062	763	451	7307	560	1354	422
HL	2000	814	779	3175	1451	859	528	7401	858	1254	394
HL	3000	897	1378	4509	1135	1079	594	9651	851	1255	368
MH	1000	2014	2511	8242	3432	1695	956	16057	1848	3367	772
MH	2000	2537	2825	8672	3449	1846	1123	18056	1917	3965	1086
MH	3000	2349	2531	9176	2681	1972	1126	16903	1617	3061	1170
MM	1000	1202	924	3855	1993	1092	580	10634	1083	1796	485
MM	2000	1108	1593	4707	1778	899	682	11460	1171	2301	800
MM	3000	1218	1729	4677	1728	1027	657	9589	967	2416	479
ML	1000	703	661	3085	1109	684	354	6111	457	1271	250
ML	2000	536	665	2960	1570	655	353	4857	726	1056	412
ML	3000	673	1335	3392	1062	750	399	5452	720	1377	391
LH	1000	1086	1893	6164	2026	1660	740	10200	1065	2031	577
LH	2000	1913	1655	5439	2277	1194	800	13562	1195	2506	872
LH	3000	1742	1415	5618	1476	1409	782	11354	928	2171	738
LM	1000	638	605	3452	1593	1154	454	6752	833	1267	306
LM	2000	1355	968	2644	1812	605	492	8346	859	1652	577
LM	3000	1226	1134	3945	787	834	296	7009	597	1189	650
LL	1000	460	599	2573	851	695	212	3729	441	777	211
LL	2000	637	577	1497	727	367	269	3858	318	854	267
LL	3000	841	330	2973	480	302	278	4630	236	796	266

Table 4: Table of manually validated calls by species for simulation 2.
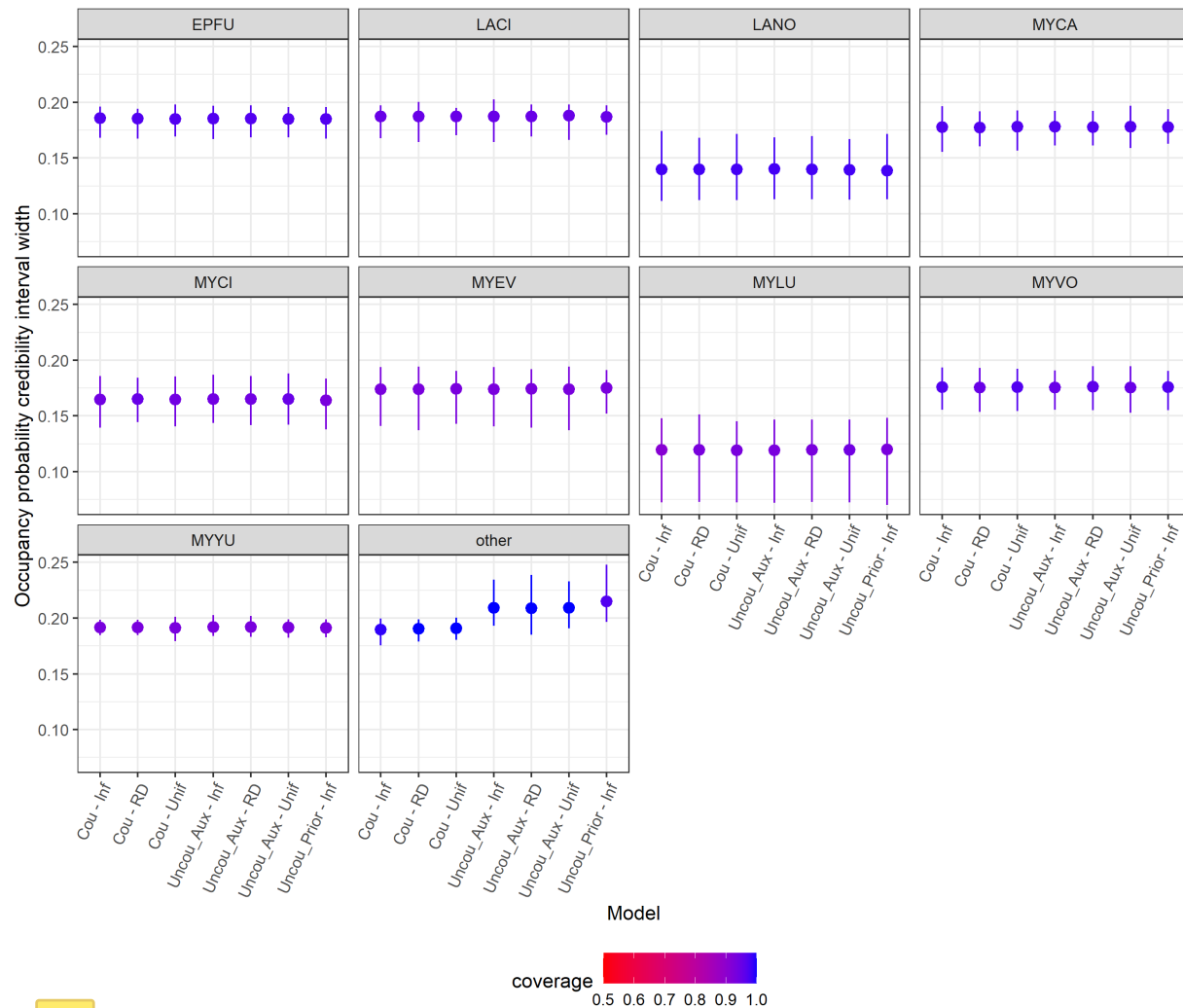
Figure 9: Credibility interval width of 95% credibility intervals for each occupancy probability. The color of each interval is determined by the proportion of simulated data sets for which the 95% credibility intervals captured the generating values. In general, interval width was similar across all considered models.

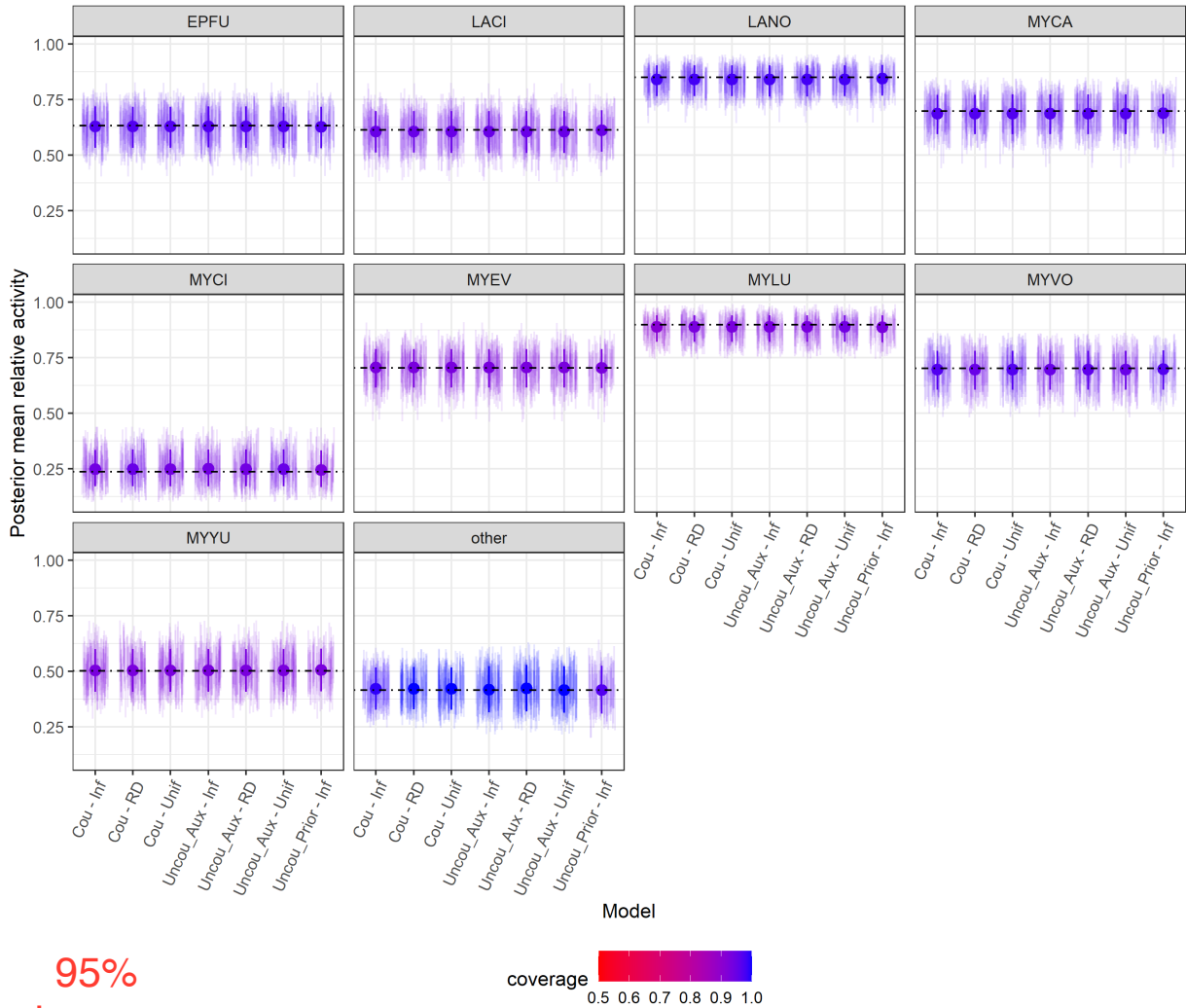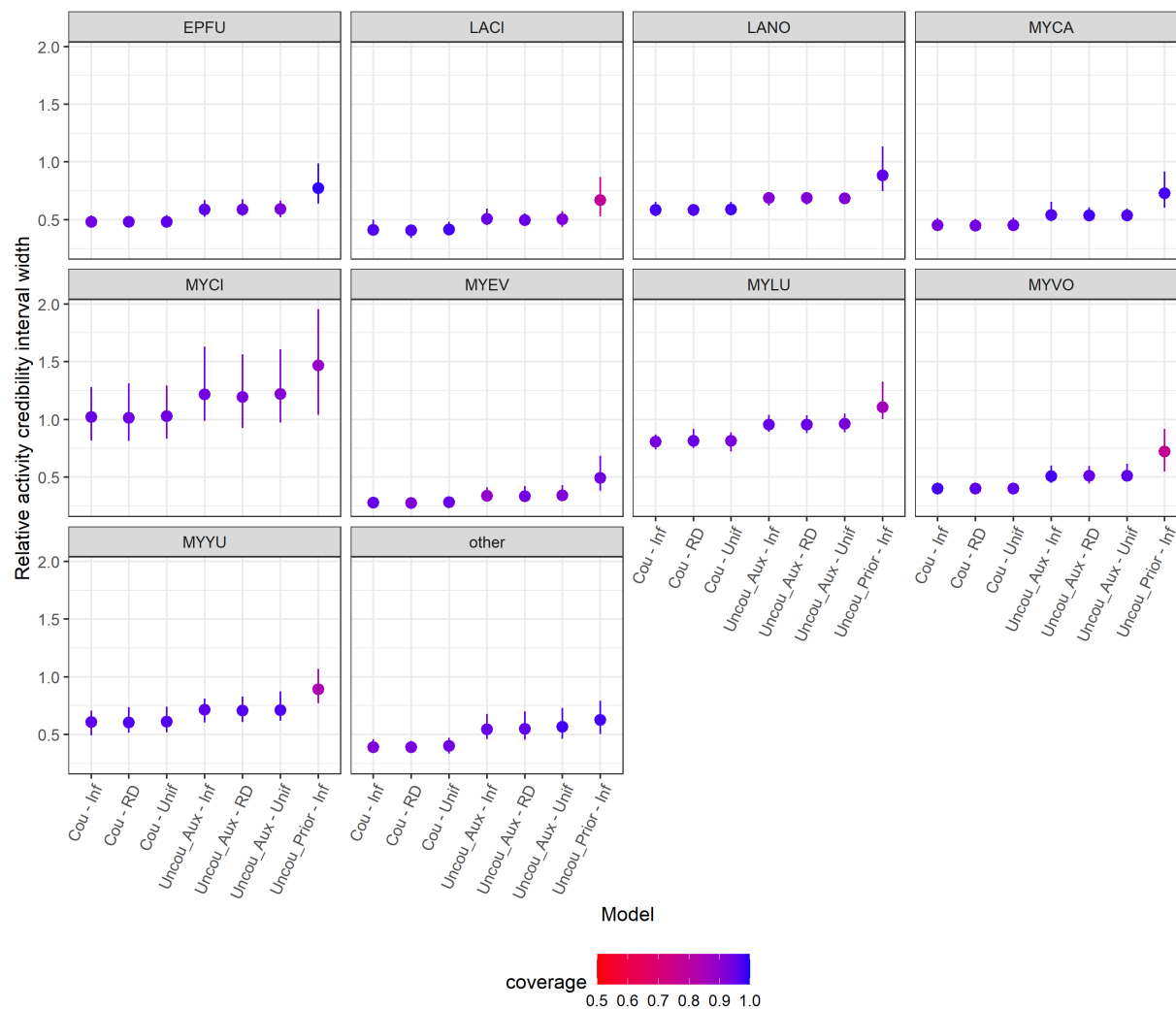*Impacts of coupling on modeling of acoustic data*
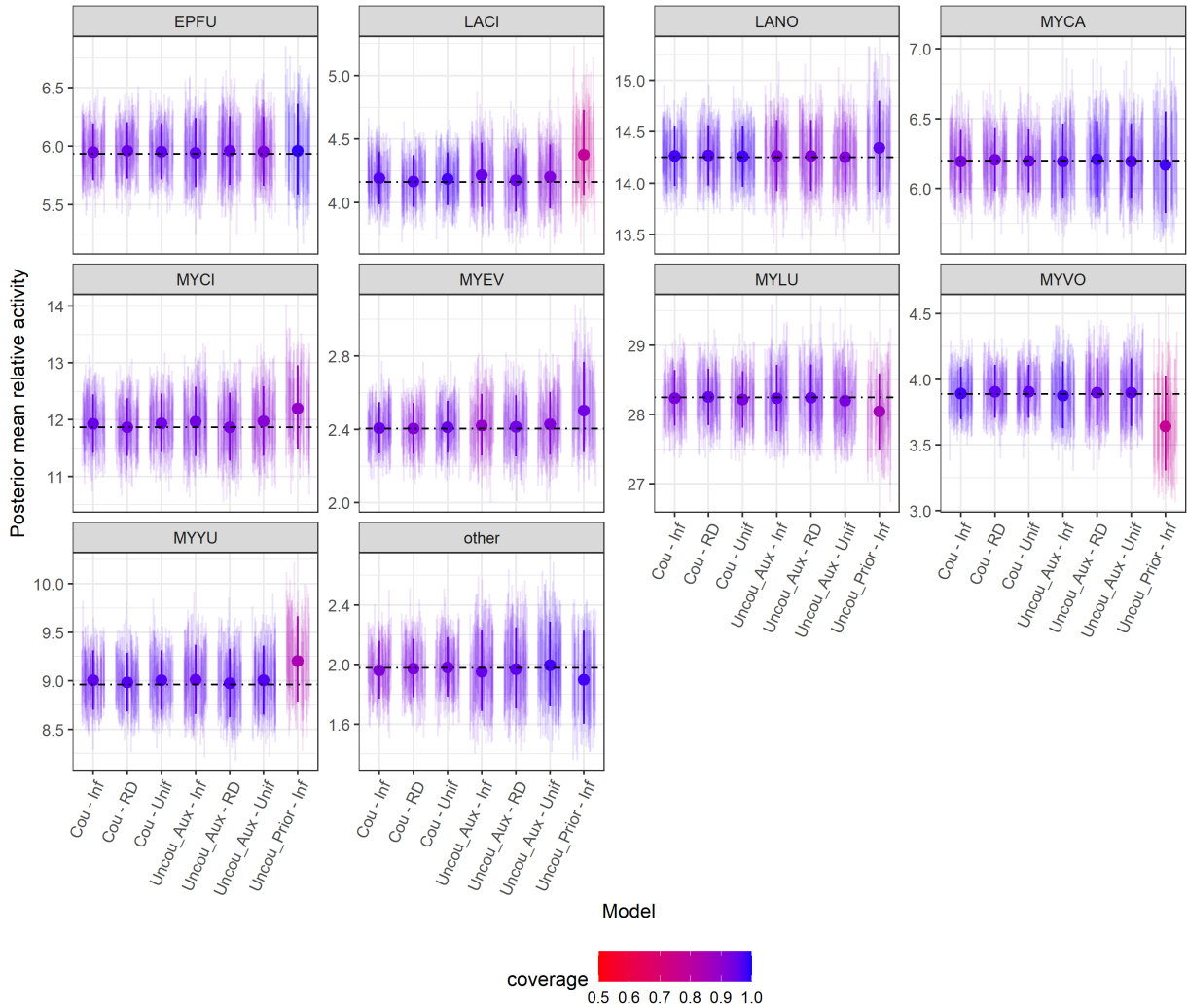

Figure 10: Credibility intervals for occupancy probability across all 100 simulations; average credibility intervals are displayed in bold. Color is determined by the proportion of simulated data sets for which credibility intervals captured the generating values.
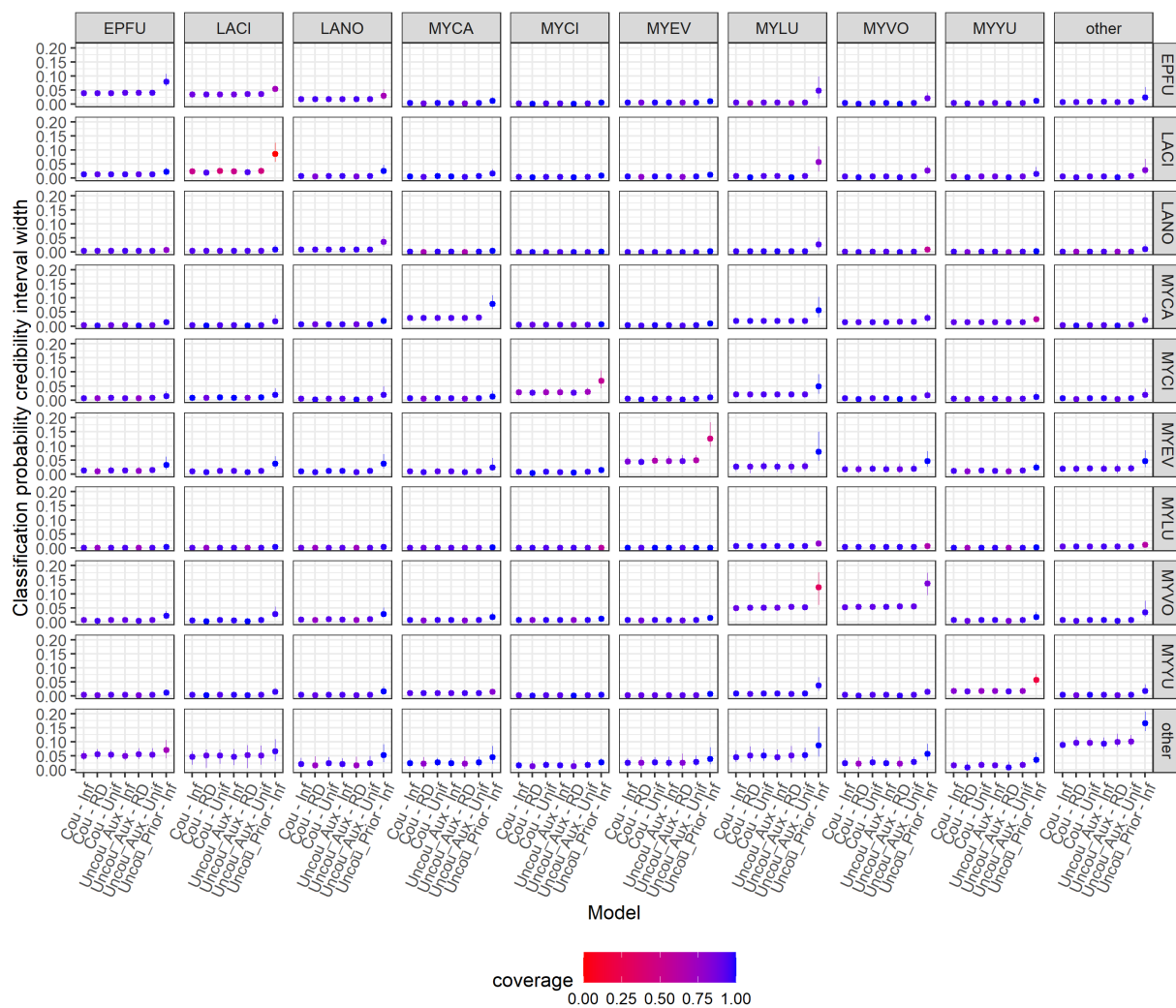
48

Impacts of coupling on modeling of acoustic data

Figure 11: Credibility interval width of 95% credibility intervals for each relative activity parameter. Color is determined by the proportion of simulated data sets for which credibility intervals captured the generating values. On average, the coupled versions of the model resulted in greater precision than the uncoupled versions. Additionally, only the reference distance prior resulted in unbiased parameter estimates and nominal coverage for all models considered.
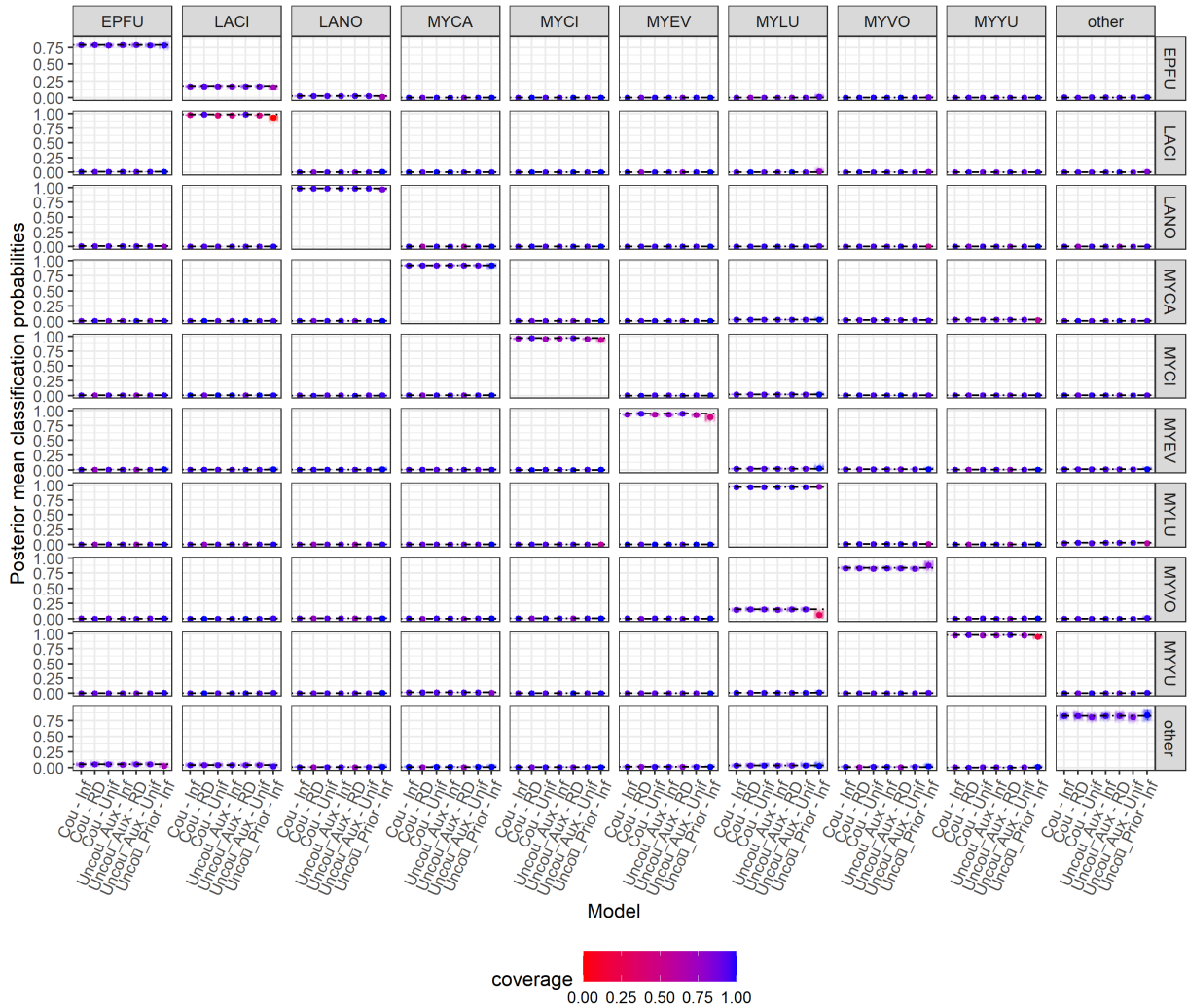
49

Figure 12: Credibility intervals for relative activity parameter across all 100 simulations; average credibility intervals are displayed in bold. Color is determined by the proportion of simulated data sets for which credibility intervals captured the generating values.

50

Impacts of coupling on modeling of acoustic data

Figure 13: Credibility interval width of 95% credibility intervals for each classification probability. The coupled model and uncoupled model with auxiliary data resulted in greater precision when estimating classification probabilities. Across all models, on average, the reference distance prior structure resulted in the least bias and greatest coverage.

Impacts of coupling on modeling of acoustic data

Figure 14: Credibility intervals for relative activity parameter across all 100 simulations; average credibility intervals are displayed in bold. Color is determined by the proportion of simulated data sets for which credibility intervals captured the generating values.
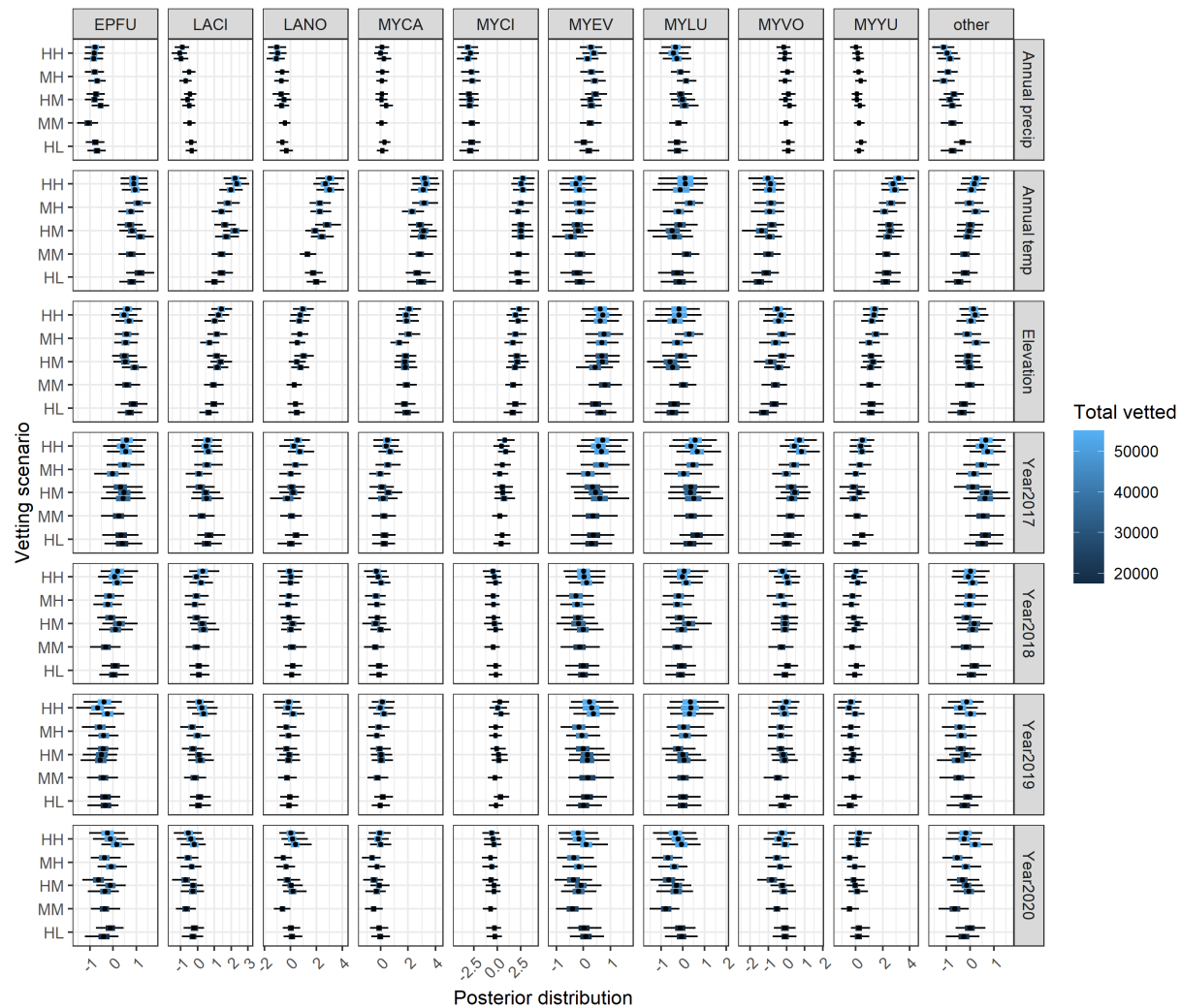
52

Figure 15: Posterior intervals of occupancy probability coefficients; points refer to the mean of the posterior distribution, thick lines to 50% credibility intervals, and thin lines to 95% credibility intervals. Vetting scenarios are arranged from top to bottom by vetting effort. Each scenario was run using three different seeds for randomization; missing intervals are indicative of lack of convergence.

53
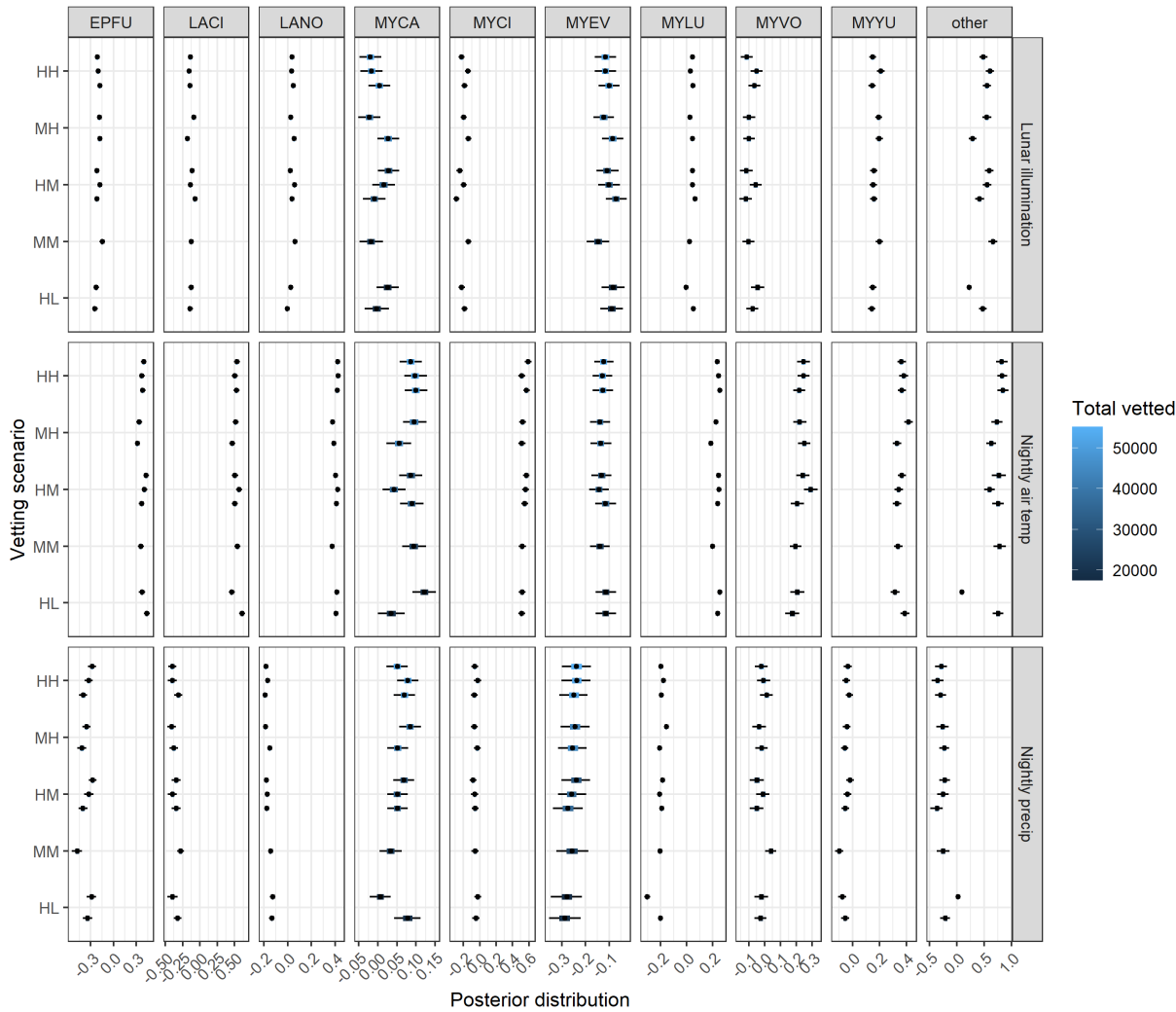
Impacts of coupling on modeling of acoustic data

Figure 16: Posterior intervals of relative activity coefficients; points refer to the mean of the posterior distribution, thick lines to 50% credibility intervals, and thin lines to 95% credibility intervals. Vetting scenarios are arranged from top to bottom by vetting effort. Each scenario was run using three different seeds for randomization; missing intervals are indicative of lack of convergence.
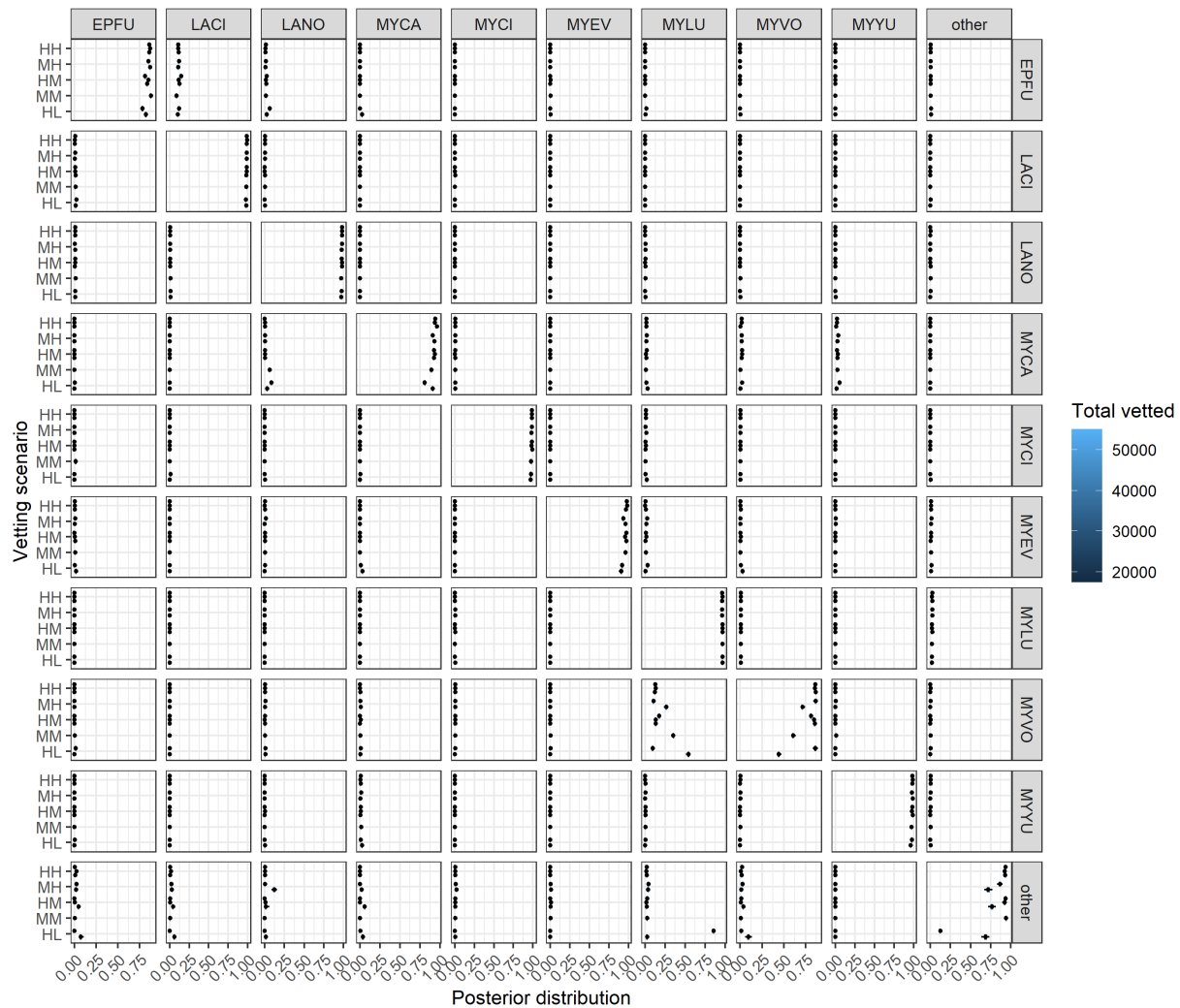
54

Impacts of coupling on modeling of acoustic data

Figure 17: Posterior intervals of classification probabilities; points refer to the mean of the posterior distribution, thick lines to 50% credibility intervals, and thin lines to 95% credibility intervals. Vetting scenarios are arranged from top to bottom by vetting effort. Each scenario was run using three different seeds for randomization; missing intervals are indicative of lack of convergence.

55